

DTIC
AD-A252 675



①

Technical Report 952

6

Building and Retaining the Career Force: New Procedures for Accessing and Assigning Army Enlisted Personnel

Annual Report, 1990 Fiscal Year

John P. Campbell and Lola M. Zook, Editors
Human Resources Research Organization

May 1992

92-17506



**United States Army Research Institute
for the Behavioral and Social Sciences**

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

**A Field Operating Agency Under the Jurisdiction
of the Deputy Chief of Staff for Personnel**

EDGAR M. JOHNSON
Technical Director

MICHAEL D. SHALER
COL, AR
Commanding

Research accomplished under contract
for the Department of the Army

Human Resources Research Organization

Technical review by

Michael G. Rumsey
Jacinto M. Silva

Accession For	
PERI-POX	<input checked="checked" type="checkbox"/>
PERI-POX	<input type="checkbox"/>
Unpublished	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POX, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.



REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE 1992, May		3. REPORT TYPE AND DATES COVERED Interim Jul 89 - Sep 90	
4. TITLE AND SUBTITLE Building and Retaining the Career Force: New Procedures for Accessing and Assigning Army Enlisted Personnel. Annual Report, 1990 Fiscal Year				5. FUNDING NUMBERS MDA903-89-C-0202 63007A 792 2208 C1	
6. AUTHOR(S) Campbell, John P., and Zook, Lola M., Editors					
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Human Resources Research Organization 66 Canal Center Plaza, Suite 400 Alexandria, VA 22314				8. PERFORMING ORGANIZATION REPORT NUMBER FR-PRD-90-6	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue Alexandria, VA 22333				10. SPONSORING/MONITORING AGENCY REPORT NUMBER ARI Technical Report 952	
11. SUPPLEMENTARY NOTES Prepared under Project Building the Career Force (Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, U.S. Army Research Institute). Contracting Officer's Representative, Michael Rumsey.					
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.				12b. DISTRIBUTION CODE --	
13. ABSTRACT (Maximum 200 words) The Career Force research project is the second phase of a two-phase U.S. Army program to develop a selection and classification system based on expected future performance for enlisted personnel. In the first phase, Project A, a large and versatile data base was collected from a representative sample of military occupational specialties and used to (a) validate the Armed Services Vocational Aptitude Battery and (b) develop and validate new predictor and criterion measures representing the entire domain of potential measures. Building on this foundation, Career Force research will finish developing the selection/classification system and evaluate its effectiveness, with emphasis on assessing second-tour performance. This first year of the project was devoted to analyzing predictor data and second-tour performance data and developing an initial model of second-tour performance.					
14. SUBJECT TERMS Career force Criterion measures Longitudinal validation Personnel classification Personnel selection Predictor measures Project A Second-tour performance				15. NUMBER OF PAGES 425	
				16. PRICE CODE --	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT Unlimited		

Technical Report 952

**Building and Retaining the Career Force:
New Procedures for Accessing and
Assigning Army Enlisted Personnel**

Annual Report, 1990 Fiscal Year

John P. Campbell and Lola M. Zook, Editors
Human Resources Research Organization

Selection and Classification Technical Area
Michael G. Rumsey, Chief

Manpower and Personnel Research Division
Zita M. Simutis, Director

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

May 1992

Army Project Number
2Q263007A792

Manpower and Personnel

Approved for public release; distribution is unlimited.

FOREWORD

This document describes the research activities conducted during the first year of the project entitled Building the Career Force. This project is the second phase of a research program of unprecedented scope and depth that will provide the basis for improving the Army's selection and classification procedures, as well as for improving reenlistment and promotion decisions for the soldiers up to the level of sergeant. The thrust for this program came from the practical, professional, and legal need to validate the Armed Services Vocational Aptitude Battery (ASVAB--the current U.S. military selection/classification test battery) and other selection variables as predictors of training and performance. The authorization for the program was provided in a letter--Deputy Chief of Staff for Operations, "Army Research Project to Validate the Predictive Value of the Armed Services Vocational Aptitude Battery," effective 19 November 1980--and a memorandum--Assistant Secretary of Defense, Manpower Reserve Affairs and Logistics, "Enlistment Standards," effective 11 September 1980.

The research program began in 1982 with an effort known as Project A. Project A not only validated the ASVAB against job performance; it further linked indicators of temperament (achievement, discipline, stress tolerance), psychomotor ability (e.g., eye-hand coordination), and spatial ability to job performance. Project A developed new tools for a variety of personnel decisions. Before these tools can be optimally used, however, two critical questions need to be answered: (1) What combinations of aptitude, temperament, psychomotor ability, and spatial ability, measured at or before entry into the Army, best predict later performance in individual military occupational specialties? (2) Which indicators of first-tour performance best predict performance in the second tour? These questions will be answered in Building the Career Force.

The first-year activities described in this report include analyses focused on the combined set of initial entry predictor measures developed for selection and classification purposes and end-of-training measures to be linked to these predictor measures. Other activities reported include analyses of the preliminary second-tour measures that will be refined and administered to a sample already tested on initial entry, end-of-training measures, and first-tour measures. Following the planned administration of the refined second-tour measures in 1991 and 1992, it will be possible to examine longitudinal linkages across the full set of measures, from initial entry to second tour. This will provide an information base for setting unrivaled selection, classification, reenlistment, and promotion policies.

This effort has been sponsored by Brigadier General Theodore G. Stroup, Director of Military Personnel Management (DMPM). Brigadier General Stroup has been briefed on the DMPM activities described in this report and has personally taken part in execution of this project in an

extremely effective manner. An Action Officers' Working Group, composed of representatives from the offices of the Director of Military Personnel Management, Deputy Chief of Staff for Personnel; Training and Doctrine Command; Forces Command; Deputy Chief of Staff for Operations; U.S. Army, Europe; Recruiting Command; Sergeant Major of the Army; Command and General Staff College; and Soldier Support Center met in May 1990 to review project objectives and plans and to provide assistance and advice.

To ensure that Building the Career Force research achieves its full scientific potential, an advisory group of experts in personnel measurement, selection, and classification was established to provide guidance on technical aspects of the research. Members of this scientific advisory group include Philip Bobko, Lloyd Bond, Milton Hakel (chair), Lloyd Humphreys, Lawrence Johnson, Robert Linn, Mary Tenopyr, and Jay Uhlaner. This group was briefed in March 1990 on research analyses conducted to that time.



EDGAR M. JOHNSON
Technical Director

EDITORS' PREFACE

This is the first annual report for work completed as part of the Building the Career Force project. It constitutes the primary technical report of the work completed on several of the project's principal tasks. Consequently, it is a "stand-alone" document and does not refer the reader to more detailed descriptions in supplementary reports. The Career Force project extends the major work on selection and classification of Army enlisted personnel that was completed as part of Project A.

The Career Force project includes (1) a replication and extension of the Experimental Battery validities for the selection and classification of first-tour enlisted personnel; (2) validation of the Experimental Battery against end-of-training performance; (3) validation of training performance as a predictor of first-tour job performance; (4) measurement of second-tour performance; (5) validation of the Armed Services Vocational Aptitude Battery, the Experimental Battery, Advanced Individual Training (AIT) performance, and first-tour performance as predictors of second-tour performance; and (6) identification of the optimal predictor battery for selection and classification, given certain specific sets of goals and constraints.

This first technical report deals with a subset of the project's principal research tasks. After an overview of the Career Force project and its relationship to Project A in chapter one, chapter two describes the basic data base for the Longitudinal Validation predictor analysis, the end-of-training criterion sample, the Longitudinal Validation criterion sample, and the second-tour Concurrent Validation criterion sample.

Chapter three presents a detailed analysis of the distributional characteristics and covariance structure of the Experimental Predictor Battery and describes the content of the "basic predictor scores" that are derived from the battery. Chapter four describes the use of data on training performance to develop six training performance criterion scores.

Chapters five and six describe how the data from the second-tour sample were used to identify the basic set of second-tour performance criterion scores and then model the latent structure of second-tour non-commissioned officer (NCO) performance. The final chapter summarizes project plans for the next year, including the follow-up criterion data collection for people from the initial longitudinal sample who reenlisted for a second tour.

In sum, by the end of the first year of the Career Force project, the basic scores had been analyzed and a latent structure proposed for the Experimental Predictor Battery, the AIT training performance measures, and the second-tour NCO performance measures. During the second year, the first-tour performance model developed in Project A will be subjected to a confirmatory test using the Longitudinal Validation sample. After that

step is complete, all the basic validation analyses will be carried out. These will be reported in the second-year annual report.

The writing of this report was as much a collective effort as the research itself. Except for chapter one, each chapter, or chapter section, was originally drafted by the team that carried out the analysis. A certain amount of cutting, splicing, rearranging, and filling in was then contributed by the editors. The authors of the original draft of each section are indicated at the appropriate place in the table of contents.

BUILDING AND RETAINING THE CAREER FORCE: NEW PROCEDURES FOR ACCESSING AND ASSIGNING ARMY ENLISTED PERSONNEL. ANNUAL REPORT, 1990 FISCAL YEAR.

EXECUTIVE SUMMARY

Requirement:

The Building the Career Force project is the second phase of a comprehensive, long-term research program, sponsored by the Deputy Chief of Staff for Personnel, to provide the basis for improving the selection and assignment of Army enlisted personnel. In the first phase, known as Project A, existing selection measures were validated against both existing and newly developed performance criteria, and new predictive measures were developed to aid in assignment and promotion decisions. The Career Force project extends the research on measuring second-tour job performance and will examine how selection and classification tests administered before a soldier's first enlistment can, with measures of performance during that enlistment, predict performance potential for second-tour duty.

Procedure:

In Task 1, measures developed in Project A to assess the performance of second-tour soldiers are being revised and tested with the Longitudinal Validation sample first tested in Project A (these second-tour soldiers have been in the Army from 41 to 63 months). The results from these tests will be used to complete the predictive validation of the Armed Services Vocational Aptitude Battery and the Project A Experimental Predictor Battery, training success measures, and first-tour job performance tests against the criteria for successful second-tour performance.

Task 2 staff is establishing, and will manage, an integrated research data base (IRDB), processing Project A and Career Force data and merging files with related military data.

Task 3 covers all analyses to be performed under this project to develop the analytic framework needed to evaluate equations for predicting training performance, first-tour performance and attrition, reenlistment, and second-tour performance.

Findings:

Establishment of the IRDB has involved organizing and initial processing of data from tests of the Longitudinal Validation Predictor sample (total sample 50,235, tested at the time of entrance into the Army, with

38,081 soldiers having complete test data); the Longitudinal Validation End-of-Training sample (total sample 44,639, of whom 34,315 have been matched to soldiers in the Longitudinal Validation Predictor sample); the Longitudinal Validation First-Tour sample (11,266, tested on performance during the tour); and the Concurrent Validation Second-Tour sample (1,053).

Comparison of initial results from administration of the Experimental Predictor Battery to the soldiers in the Longitudinal Evaluation Predictor sample indicated no major inconsistencies with earlier results from Project A testing. A set of composite scores was developed from the Battery's paper-and-pencil and computer-administered measures to be used in later validation and prediction analyses.

A set of six factor scores was developed from the End-of-Training job knowledge and rating scores to serve as basic criteria for training performance. They correspond directly to the performance components identified earlier in modeling from Concurrent Validation first-tour data.

Data from the administration of the revised second-tour measures to the Concurrent Validation Second-Tour sample were organized to provide 22 basic criterion scores. These in turn yielded a set of six performance factors characterizing second-tour job performance. There was substantial correspondence between the first-tour and the second-tour models of performance.

Utilization of Findings:

The data processed during the first year of the Career Force project will provide bases for further analyses as collection of Longitudinal Validation data continues. The long-term results from these developmental and validation processes will be applied in an improved system for selecting and assigning Army manpower.

**BUILDING AND RETAINING THE CAREER FORCE: NEW PROCEDURES FOR ACCESSING AND
ASSIGNING ARMY ENLISTED PERSONNEL. ANNUAL REPORT, 1990 FISCAL YEAR**

CONTENTS

	Page
INTRODUCTION	1
(John P. Campbell and James Harris)	
Characteristics of Present Army Personnel System	1
Recruitment	1
Selection and Classification at the Processing Station	4
Initial Classification	5
Initial Training	6
Performance Assessment in Army Units	7
Reenlistment Screening	7
Summary	8
A Brief History of Selection and Classification	8
A Summary Description of Project A	9
Project A Task Outline	10
The Organization of Project A	11
The Research Plan and Integrated Master Plan	13
MOS and Sample Selection	17
Predictor Development	18
Performance Measurement	21
The Concurrent Validation (CV)	26
The Longitudinal Validation (LV) Data Collections	34
Second-Tour Performance Criterion Development	35
Job Performance Measurement	38
Summary	38
Building on Project A	39
The Foundation Provided by Project A	39
Project A Products and Results	40

CONTENTS (Continued)

	Page
Building the Career Force	42
Project Objectives	43
Project Organization	44
Summary	46
Content of This Report	46
DATA FILE DESIGN AND PREPARATION	49
(Diane Steele, Scott Oppler, and Winnie Young)	
Longitudinal Validation (LV) Predictor Sample Files	49
Paper-and-Pencil Data Files	52
Computer Battery Data Files	54
Longitudinal Validation End-of-Training Data Files	55
School Knowledge Test Data Files	57
End-of-Training Ratings Data Files	59
Concurrent Validation Second-Tour (CVII) Files	61
Longitudinal Validation First-Tour (LVI) Data	66
Updates from Existing Army Data	71
Generation of Career Force Project Workfiles	72
ANALYSIS OF THE EXPERIMENTAL PREDICTOR BATTERY: LV SAMPLE	73
(Norman Peterson, Teresa Russell, Glenn Hallam, Leaetta Hough, Cynthia Owens-Kurtz, Kathleen Gialluca, and Kathryn Kerwin)	
Introduction	73
The Experimental Battery	74
The Analysis Samples	74
Scoring and Forming Composites of Paper-and-Pencil Predictors	78
Descriptions of Tests and Test History	78
Evaluation of Methods for Screening Scores on the Six Spatial Tests	82
Analysis of Alternative Scores	88

CONTENTS (Continued)

	Page
Comparison of Gender and Race Subgroup Scores	91
Analyses and Conclusions Regarding Composite Formation	94
Scoring and Forming Composites of Computer-Administered Predictor Scores	101
Test Descriptions	102
Variance Analysis of Selected Computer Test Scores	102
Analysis of Possible Scoring Methods/Changes	115
Final Screening Rules	119
Basic Scores for Further Analysis	120
Comparison of Gender and Race Subgroup Scores	124
Composite Formation	130
Summary of Computer-Administered Test Score Composites	135
Comparison of Longitudinal Initial Sample and Sample 2	138
Reliability Estimates for Computer Test Composites	141
Analyses of Cognitive Predictor Composite Scores	142
Scoring and Forming Composites for the ABLE Inventory	142
Development and Content of the ABLE Inventory	146
Data Screening	146
Analyses to Verify Appropriateness of the Scoring Procedures	148
Comparison of Descriptive Statistics for the Revised Trial Battery and Experimental Battery	148
Analysis of Subgroup Differences	150
Uniqueness Analysis	150
Formation of ABLE Composites	152
Scoring and Forming Composites for the AVOICE Inventory	163
Development and Content of the AVOICE Inventory	163
Data Screening	164
Comparison of Descriptive Statistics for the Revised Trial Battery and Experimental Battery	164
Analysis of Subgroup Differences	168
Uniqueness Analysis	171
Formation of AVOICE Composites	171

CONTENTS (Continued)

	Page
Scoring and Forming Composites for the JOB Inventory	177
Development and Content of the JOB Inventory	177
Data Screening	181
Comparison of Descriptive Statistics for the Revised Trial Battery and Experimental Battery	184
Analysis of Subgroup Differences	184
Uniqueness Analyses	186
Formation of JOB Composites	186
Longitudinal Validation Predictors: Summary of Data Analyses and Formation of Composites	192
Data Screening and Basic Scoring	192
Descriptive Statistics and Psychometric Properties	193
Formation of Composite Scores	194
A Final Word	199
END-OF-TRAINING MEASURES: LV SAMPLE (Rodney A. McCloy and Scott Oppler)	201
The End-of-Training (EOT) School Knowledge (SK) Scores	201
Revision of the EOT SK Test Items and Scoring Keys	201
Analyses of the Revised EOT SK Tests	204
Creation of EOT School Knowledge Test Factor Scores	212
End-of-Training Ratings	218
Data Editing and Outlier Analyses	221
EOT Ratings Analysis Samples	222
Descriptive Statistics and Reliabilities	224
Exploratory Factor Analyses	229
Confirmatory Factor Analyses	229
Creation of EOT Rating Factor Scores	242
Relationships Between the EOT Factor Scores	244
DEVELOPMENT OF SCORES FOR SECOND-TOUR PERFORMANCE MEASURES (Walter C. Borman, Charlotte H. Campbell, Mary Ann Hanson, Jerry W. Hedge, Deirdre J. Knapp, Dennis P. McGuire, Elaine D. Pulakos, and Deborah Whetzel)	247

CONTENTS (Continued)

	Page
Development of the Second-Tour Measures	247
Modifications of First-Tour Measures for	
Second-Tour Use	247
New Measurement Methods for Second-Tour Performance	250
Supplemental Information	252
Second-Tour Data Collection	252
Data Collection Procedures	254
Sample Sizes	257
Analysis for CVII Basic Criterion Scores	258
CVII Army-Wide and MOS-Specific Rating Scales	258
Combat Performance Prediction Scales	279
Basic Scores for Second-Tour Hands-On Job	
Samples and Job Knowledge Tests	282
Basic Scores for Administrative Measures	297
The Situational Judgment Test	302
NCO Supervisory Simulation Exercises	320
Army Job Satisfaction Questionnaire (AJSQ)	330
Summary of Second-Tour Criterion Score Development	333
MODELING OF SECOND-TOUR PERFORMANCE	337
(John P. Campbell and Scott Oppler)	
The Input Data	337
Procedure	340
Results	341
Comparisons of Models	341
Subsample Differences	349
Final Scores	349
Implications	358
FUTURE CAREER FORCE PROJECT PLANS	361
(John P. Campbell and Deirdre J. Knapp)	

CONTENTS (Continued)

	Page
LVII Data Collection	361
Anticipated Problems	361
Research Support Requests	363
Revision of Performance Measures	364
Data Analysis Plans	364
Near-Term Analyses	364
Longer Term Analyses	367
REFERENCES	369
APPENDIX A. SAMPLE DESCRIPTIONS OF COGNITIVE MEASURES	A-1
B. DEFINITIONS OF ELEMENTS IN THE NON-COGNITIVE INVENTORIES	B-1

LIST OF TABLES

Table 1.1	The Army Selection, Classification, and Evaluation Process	2
1.2	Initial Project A Military Occupational Specialties (MOS)	18
1.3	Hierarchical Map of Predictor Space	20
1.4	Summary of Predictor Measures Used in Concurrent Validation (The Trial Battery)	22
1.5	Summary of Criterion Measures Used in Batch A and Batch Z Concurrent Validation Samples	25
1.6	Concurrent Validation Sample Soldiers by MOS by Location	27
1.7	Predictor Construct Scores From Concurrent Validation Data	28
1.8	Latent Structure Scores From Concurrent Validation Data	29
1.9	Mean Validity for the Composite Scores Within Each Predictor Domain Across Nine Army Enlisted Jobs	31

CONTENTS (Continued)

	Page
Table 1.10 Mean Incremental Validity for the Composite Scores Within Each Predictor Domain Across Nine Army Enlisted Jobs	32
1.11 Description of Tests in the Experimental Predictor Battery	36
2.1 Longitudinal Validation (LV) Predictor Sample by MOS: Total Sample	50
2.2 LV Predictor Sample by Gender: Total Sample	50
2.3 LV Predictor Sample by Race: Total Sample	51
2.4 LV Predictor Sample by MOS: Sample With Both Computer and Paper-and-Pencil Batteries	51
2.5 LV Predictor Sample by Gender: Sample With Both Computer and Paper-and-Pencil Batteries	52
2.6 LV Predictor Sample by Race: Sample With Both Computer and Paper-and-Pencil Batteries	52
2.7 LV End-of-Training (EOT) Sample by MOS: Cases Matched With LV Predictor Sample	56
2.8 LV End-of-Training (EOT) Sample by Gender: Cases Matched With LV Predictor Sample	56
2.9 LV End-of-Training Sample by Race: Cases Matched With LV Predictor Sample	57
2.10 LV End-of-Training Rater-Ratee Sample by MOS	60
2.11 LV End-of-Training Rater-Ratee Sample by Gender	60
2.12 LV End-of-Training Rater-Ratee Sample by Race	61
2.13 Concurrent Validation (CVII) Sample by MOS	62
2.14 CVII Sample by Gender	62
2.15 CVII Sample by Race	62
2.16 LVI Sample by MOS	67

CONTENTS (Continued)

	Page
Table 2.17 LVI Sample by Gender	67
2.18 LVI Sample by Race	68
3.1 Longitudinal Validation: Comparison of Initial Sample and Sample 2 Demographics to the Complete Sample by MOS	76
3.2 Longitudinal Validation: Comparison of Initial Sample and Sample 2 Demographics to the Complete Sample by Race and Gender	77
3.3 Changes in Cognitive Paper-and-Pencil Measures From Pilot Trial Battery to Trial Battery (Concurrent) to Experimental Battery (Longitudinal)	79
3.4 Cognitive Paper-and-Pencil Measures Comple- tion Rates and Ceiling and Floor Effects: Concurrent Validation and Initial Longitudinal Samples	80
3.5 Cognitive Paper-and-Pencil Measures: Reliability Comparisons Between Pilot Trial Battery, Trial Battery, and Experimental Battery Administrations . .	81
3.6 Cognitive Paper-and-Pencil Measures: Comparison of Correlations of Number Correct Score in Concur- rent and Longitudinal Validations	83
3.7 Accuracy Score Statistics Used to Define Suspect Score Ranges on Six Paper-and-Pencil Tests	85
3.8 Longitudinal Validation Modified Caution Index: Descriptive Statistics on Six Paper-and-Pencil Tests for Initial Sample	85
3.9 Subjects Flagged by Applying Accuracy by Modified Caution Index Criteria on Six Paper-and-Pencil Tests: Initial Longitudinal Sample	86
3.10 AFQT Means for the Initial Longitudinal Sample and Examinees Flagged by the Modified Caution Index by Accuracy Criteria	86

CONTENTS (Continued)

	Page
Table 3.11 Longitudinal Validation Runs Test: Descriptive Statistics on Six Paper-and-Pencil Tests for Initial Sample	87
3.12 Subjects Flagged by Applying Accuracy by Runs Test Criteria on Six Paper-and-Pencil Tests: Initial Longitudinal Sample	87
3.13 Longitudinal Validation: Five Alternative Scores on Cognitive Paper-and-Pencil Measures for Initial Sample	89
3.14 Longitudinal Validation: Squared Multiple Regression Coefficients, Reliability Estimates, and Uniqueness Estimates for Cognitive Paper-and-Pencil Measures for Initial Sample	90
3.15 Cognitive Paper-and-Pencil Measures: Means and Effect Sizes for Number Correct by Gender (CV Sample, LV Initial Sample, and LV Sample 2)	92
3.16 Cognitive Paper-and-Pencil Measures: Means and Effect Sizes for Number Correct by Race (CV Sample, LV Initial Sample, and LV Sample 2)	93
3.17 Cognitive Paper-and-Pencil Measures: Factor Loadings for Number Correct, Principal Factor Analyses Two-Factor Solution (CV Sample and LV Initial Sample)	95
3.18 Concurrent Validation Cognitive Paper-and-Pencil Measures: Factor Loadings for Number Correct and ASVAB Subtests, Principal Factor Analysis Five-Factor Solution	96
3.19 Longitudinal Validation Cognitive Paper-and-Pencil Measures: Factor Loadings for Number Correct and ASVAB Subtests, Principal Factor Analysis Five-Factor Solution (Initial Sample)	97
3.20 LISREL Runs on Initial Longitudinal Sample Spatial Test Data to Examine Four Alternate Composite Models	98
3.21 Second-Order Analysis of Spatial Test Scores: Schmid-Leiman Transformation	100

CONTENTS (Continued)

	Page
Table 3.22 Comparison of Cognitive Paper-and-Pencil Test Factor Loadings for Two Longitudinal Validation Samples	101
3.23 Longitudinal Validation: Changes in Computer- Administered Measures From Pilot Trial Battery to Trial Battery to Experimental Battery	103
3.24 Concurrent and Longitudinal Validations: Mean Time to Read Instructions and Complete Test Items for Computer-Administered Tests	104
3.25 Computer-Administered Tests: Analysis of Variance Due to Item Parameters--Perceptual Speed and Accuracy	107
3.26 Computer-Administered Tests: Analysis of Variance Due to Item Parameters--Short-Term Memory	108
3.27 Computer-Administered Tests: Analysis of Variance Due to Item Parameters--Target Tracking Test 1	109
3.28 Computer-Administered Tests: Analysis of Variance Due to Item Parameters--Target Tracking Test 2	109
3.29 Computer-Administered Tests: Analysis of Variance Due to Item Parameters--Target Shoot Test	110
3.30 Computer-Administered Tests: Analysis of Variance Due to Item Parameters--Cannon Shoot Test	111
3.31 Computer-Administered Tests: Analysis of Variance Due to Item Parameters--Target Identification Test	112
3.32 Computer-Administered Tests: Analysis of Variance Due to Item Parameters--Number Memory Operation Time	112

CONTENTS (Continued)

	Page
Table 3.33 Computer-Administered Reaction Time Tests: Reliability Coefficients, Squared Multiple Regression Coefficients (vs. All ASVAB Subtests), and Uniqueness Estimates	117
3.34 Alternate Scores for Target Shoot Test: Reliability and Squared Multiple Regression Coefficients (vs. All ASVAB Subtests), and Uniqueness Estimates	118
3.35 Means of Computer-Administered Cognitive/ Perceptual Measures From Concurrent Vali- dation and Longitudinal Validation Samples	121
3.36 Reliability Estimates for Computer-Administered Cognitive/Perceptual Test Scores	122
3.37 Means of Computer-Administered Psychomotor Tests from Concurrent Validation and Longi- tudinal Validation Samples	123
3.38 Reliability and Uniqueness Estimates for Computer-Administered Psychomotor Test Scores	123
3.39 Computer-Administered Tests: Means and Effect Sizes by Gender (Concurrent, Longitudinal Initial, and Longitudinal Sample 2 Samples)	125
3.40 Computer-Administered Tests: Means and Effect Sizes by Race (Concurrent, Longitudinal Initial, and Longitudinal Sample 2 Samples)	127
3.41 Concurrent Validation: Factor Analysis of Computer-Administered Measures	130
3.42 Longitudinal Validation Initial Sample: Factor Analysis of Computer-Administered Measures	131
3.43 Concurrent Validation: Factor Analysis of Computer-Administered Measures and ASVAB Subtests	133
3.44 Longitudinal Validation Initial Sample: Factor Analysis of Computer-Administered Tests and ASVAB Subtests	134

CONTENTS (Continued)

	Page
Table 3.45 Longitudinal Validation: Correlations Between Proportion Correct and Time Scores for Total and Alternate Half Scores on Per- ceptual Computer-Administered Tests	138
3.46 Principal Factor Analysis of Total Scores for 33 Cognitive Predictors: Longitudinal Initial Sample	139
3.47 Principal Factor Analysis of Total Scores for 33 Cognitive Predictors: Longitudinal Sample 2	140
3.48 Initial Consistency Estimates for Computer Test Composites: Longitudinal Initial Sample and Sample 2	141
3.49 Correlations Between Cognitive Composites: Longitudinal Initial Sample and Sample 2	143
3.50 Cognitive Composite Score Means and Effect Sizes by Gender for Longitudinal Initial Sample and Sample 2	144
3.51 Cognitive Composite Score Means and Effect Sizes by Race for Longitudinal Initial Sample and Sample 2	145
3.52 Longitudinal Validation: Option Endorsement Rates for Items on the ABLE Non-Random Response Scale	147
3.53 Comparison of CV and LV ABLE Data Screening Results	147
3.54 Comparison of ABLE Scale Scores and Reliabilities from the Revised Trial (CV) and Experimental (LV) Batteries	149
3.55 Longitudinal Validation: ABLE Scale Score Means and Effect Sizes by Gender	150
3.56 Longitudinal Validation: ABLE Scale Score Means and Effect Sizes by Race	151

CONTENTS (Continued)

	Page
Table 3.57 Comparison of Reliability Coefficients, Multiple Regression Coefficients, and Uniqueness Estimates for ABLE Scale Scores: CV/LV	152
3.58 Comparison of ABLE Scale Intercorrelations from the Revised Trial (CV) and Experimental (LV) Batteries: CV/LV	154
3.59 Principal Components Analysis of the ABLE Content Scales: Initial Longitudinal Sample	155
3.60 Principal Factor Analysis of the ABLE Content Scales: Initial Longitudinal Sample	156
3.61 Comparison of ABLE Principal Components Analysis Results: CV/LV	157
3.62 Longitudinal Validation: Comparison of Four ABLE Composite Formation Models, Using LISREL	158
3.63 Longitudinal Validation: ABLE Composite Score Reliability and Uniqueness Estimates	160
3.64 Longitudinal Validation: Correlations Among ABLE Composites	161
3.65 Longitudinal Validation: ABLE Composite Score Means and Effect Sizes by Gender	161
3.66 Longitudinal Validation: ABLE Composite Score Means and Effect Sizes by Race	162
3.67 Frequency Distribution of Scores on the AVOICE Chi-Square Patterned Response Index (With Cut Score)	166
3.68 Frequency Distribution of Scores on the AVOICE Runs Index (With Cut Score)	166
3.69 Frequency Distribution of Scores on the AVOICE Option Variance Index (With Cut Score)	167
3.70 Frequency Distribution of Scores on the AVOICE Unlikely Response Scale (With Cut Score)	167

CONTENTS (Continued)

	Page
Table 3.71 Comparison of CV and LV AVOICE Data Screening Results	168
3.72 Comparison of AVOICE Scales Scores and Reliabilities for the Revised Trial (CV) and Experimental (LV) Batteries	169
3.73 Longitudinal Validation: AVOICE Scale Score Means and Effect Sizes by Gender	170
3.74 Longitudinal Validation: AVOICE Scale Score Means and Effect Sizes by Race	172
3.75 Comparison of Reliability Coefficients, Multiple Regression Coefficients, and Uniqueness Estimates for AVOICE Scale Scores: CV/LV	173
3.76 Longitudinal Validation: Correlations Among 22 AVOICE Scale Scores	174
3.77 Longitudinal Validation: Comparison of Four AVOICE Composite Formation Models, Using LISREL . . .	175
3.78 Longitudinal Validation: AVOICE Composite Score Reliability and Uniqueness Estimates	178
3.79 Longitudinal Validation: Correlations Among AVOICE Composites	178
3.80 Longitudinal Validation: AVOICE Composite Score Means and Effect Sizes by Gender	179
3.81 Longitudinal Validation: AVOICE Composite Score Means and Effect Sizes by Race	180
3.82 Longitudinal Validation: Frequency Distribution of Scores on the JOB Chi-Square Patterned Response Index (With Cut Score)	182
3.83 Longitudinal Validation: Frequency Distribution of Scores on the JOB Runs Test (With Cut Score) . . .	182
3.84 Longitudinal Validation: Frequency Distribution of Scores on the JOB Option Variance Index (With Cut Score)	183

CONTENTS (Continued)

	Page
Table 3.85 Longitudinal Validation: Frequency Distribution of Scores on the JOB Unlikely Response Scale (With Cut Score)	183
3.86 Comparison of CV and LV JOB Data Screening Results	184
3.87 Comparison of JOB Scale Scores and Reliabilities for Revised Trial (CV) and Experimental (LV) Batteries	185
3.88 Longitudinal Validation: JOB Score Means and Effect Sizes by Gender	186
3.89 Longitudinal Validation: JOB Score Means and Effect Sizes by Race	187
3.90 Longitudinal Validation: Reliability Coefficients, Multiple Regression Coefficients, and Uniqueness Estimates for JOB Scale Scores: CV/LV	188
3.91 Comparison of JOB Scale Intercorrelations for Revised Trial and Experimental Batteries: CV/LV . . .	188
3.92 Comparison of JOB Principal Components Analysis Results for Revised Trial (CV) and Experimental Batteries: CV/LV	189
3.93 Longitudinal Validation: Correlations Among JOB Composite Scores	190
3.94 Longitudinal Validation: JOB Composite Score Reliability and Uniqueness	190
3.95 Longitudinal Validation: JOB Composite Score Means and Effect Sizes by Gender	191
3.96 Longitudinal Validation: JOB Composite Score Means and Effect Sizes by Race	191
3.97 Experimental Battery Scores for Longitudinal Validation Initial Sample: Reliabilities, Uniqueness Estimates, and Effect Size	196

CONTENTS (Continued)

	Page
Table 3.98 Correlations Between Experimental Battery Composite Scores for Longitudinal Validation Initial Sample	197
4.1 Summary of Item Alterations for the EOT School Knowledge Tests	205
4.2 Functional Categories for the EOT School Knowledge Tests Along With Their CVBITS Classification	206
4.3 CVBITS Factors Included in the Confirmatory Factor Analyses of Model CFA3	208
4.4 Number of Soldiers Included in the Confirmatory Factor Analyses and the Correlational Analyses	208
4.5 Indexes of Fit Generated by Various Models for the EOT SK Tests	210
4.6 Functional Categories Comprising the Two Subscores for the EOT School Knowledge Tests	213
4.7 Mean Peer and Supervisor EOT Ratings Per Ratee, by MOS: Unedited Data	220
4.8 Mean Peer and Supervisor EOT Ratings Per Ratee, by MOS: Edited Data	223
4.9 Number of Ratees With a Minimum of Two Complete Sets of Peer EOT Ratings and One Complete Set of Supervisor EOT Ratings, by MOS	225
4.10 Number of Ratees With a Minimum of Two Complete Sets of Supervisor EOT Ratings, by MOS	225
4.11 Means of EOT Rating Scales: Primary Analysis Sample	226
4.12 Single-Rater/Two-Rater Reliabilities of EOT Rating Scales: Primary Analysis Sample	228
4.13 Single-Rater/Two-Rater Reliabilities of EOT Rating Scales: Supervisor Ratings Reliability Sample	228

CONTENTS (Continued)

	Page
Table 4.14 Exploratory Factor Analysis of EOT Rating Scales: Proportion of Common Variance Accounted for by First Seven Unrotated Factors	230
4.15 Exploratory Factor Analysis of EOT Rating Scales: Factor Loadings for One-Factor Solutions	231
4.16 Confirmatory Factor Analysis: Comparison of Fit of One- and Four-Factor Models, Based on Peer 1 EOT Ratings	232
4.17 Confirmatory Factor Analysis: Pattern Matrix and Factor Correlation Matrix for Four-Factor Model, Based on Peer 1 EOT Ratings	233
4.18 Confirmatory Factor Analysis: Fit of One-Factor Model, Based on Peer 1 EOT Ratings, by MOS	234
4.19 Confirmatory Factor Analysis: Fit of Four-Factor Model, Based on Peer 1 EOT Ratings, by MOS	235
4.20 Confirmatory Factor Analysis: Comparison of Separately and Simultaneously Estimated Solutions of Four-Factor Model for Peer 1 and Peer 2 EOT Ratings	236
4.21 Confirmatory Factor Analysis: Comparison of Fit of One- and Four-Factor Models, Based on Supervisor EOT Ratings	237
4.22 Confirmatory Factor Analysis: Pattern Matrix and Factor Correlation Matrix for Four-Factor Model, Based on Supervisor EOT Ratings	238
4.23 Confirmatory Factor Analysis: Fit of One-Factor Model, Based on Supervisor EOT Ratings, by MOS	239
4.24 Confirmatory Factor Analysis: Fit of Four-Factor Model, Based on Supervisor EOT Ratings, by MOS	240

CONTENTS (Continued)

	Page
Table 4.25 Confirmatory Factor Analysis: Comparison of Separately and Simultaneously Estimated Solutions of Four-Factor Model for Peer 1 and Supervisor Ratings	241
4.26 Means of EOT Rating Factor Scores: Primary Analysis Sample	243
4.27 Single-Rater/Two-Rater Reliabilities of EOT Rating Factor Scores: Primary Analysis Sample	244
4.28 Average Correlations of EOT SK Test Scores with EOT Rating Scores Across Seven Batch A MOS Peer Raters	245
4.29 Average Correlations of EOT SK Test Scores With EOT Rating Scores Across Two Batch A MOS Peer Raters	245
4.30 Average Correlations of EOT SK Test Scores With EOT Rating Scores Across Seven Batch A MOS Supervisor Raters	246
4.31 Average Correlations of EOT SK Test Scores With EOT Rating Scores Across Two Batch A MOS Supervisor Raters	246
5.1 Second-Tour Criterion Measures and Supplemental Information	253
5.2 CVII Data Collection Test Dates, 1988-89	253
5.3 CVII Data Collection Totals	257
5.4 Number of CVII Peer and Supervisor Ratings Per Ratee by MOS	260
5.5 CVII Army-Wide Ratings: Use of Scale Points by Peers	262
5.6 CVII Army-Wide Ratings: Use of Scale Points by Supervisors	263
5.7 CVII Army-Wide Ratings: Means and Standard Deviations	264

CONTENTS (Continued)

	Page
Table 5.8 CVII Army-Wide Ratings: One-Rater Interrater Reliability	265
5.9 CVII Army-Wide Ratings: <u>k</u> -Rater Interrater Reliability	266
5.10 Comparison of CVI and CVII Factor Analyses: Pooled Peer/Supervisor Ratings, Non-Supervisory Dimensions Only	268
5.11 CVII Army-Wide Factor Analysis: Peer Ratings, All Dimensions	269
5.12 CVII Army-Wide Factor Analysis: Supervisor Ratings, All Dimensions	270
5.13 CVII Army-Wide Factor Analysis: Pooled Peer/ Supervisor Ratings, All Dimensions	271
5.14 CVII MOS-Specific Ratings: Means Across Rating Dimensions for Each MOS	273
5.15 CVII MOS-Specific Ratings: One-Rater Interrater Reliability by MOS	274
5.16 CVII MOS-Specific Ratings: <u>k</u> -Rater Interrater Reliability by MOS	275
5.17 Composition of Proposed CVII Army-Wide Rating Composites	276
5.18 Definitions of Proposed CVII Army-Wide Rating Composites	277
5.19 One-Rater and <u>k</u> -Rater Interrater Reliability for Army-Wide Composites	278
5.20 Intercorrelations Among Proposed CVII Rating Composites for Pooled Peer/Supervisor Ratings	279
5.21 Principal Components Analysis of Combat Perform- ance Prediction Scale Ratings	281
5.22 Combat Performance Prediction Scales Interrater Reliability Estimates	281

CONTENTS (Continued)

	Page
Table 5.23 Combat Performance Prediction Scale Descriptive Statistics, by MOS	282
5.24 Revisions to Job Knowledge Components to Eliminate Marginal Items	283
5.25 Number of Tasks Dropped From Hands-On Component Due to Missing Data	284
5.26 Extent of Missing Data on Hands-On and Job Knowledge Components	286
5.27 Number of Hands-On Task Tests and Steps and Number of Job Knowledge Task Tests and Items for Nine MOS (Second Tour)	287
5.28 Statistical Characteristics of Hands-On and Job Knowledge Components for Each MOS	288
5.29 MOS Task Representation in Functional Categories, Task Factors, and Task Constructs (Number of Job Knowledge Tests/Number of Hands-On Tests)	292
5.30 Statistical Characteristics of Functional Categories for Hands-On and Job Knowledge Components for Nine MOS	293
5.31 Statistical Characteristics of Task Factors for Hands-On and Job Knowledge Components for Nine MOS	295
5.32 Statistical Characteristics of Task Constructs for Hands-On and Job Knowledge Components for Nine MOS	296
5.33 Statistical Characteristics of Hands-On and Job Knowledge Component Basic Task Scores and Technical Task Scores Across Nine MOS	298
5.34 Second-Tour Personnel File Form Score Means and Standard Deviations	301
5.35 Second-Tour Personnel File Form Score Intercorrelations	302

CONTENTS (Continued)

	Page
Table 5.36 Situational Judgment Test: Supervisory Behavior Dimension Definitions	307
5.37 Situational Judgment Test: Means and Internal Reliabilities	309
5.38 Situational Judgment Test: Score Intercorrela- tions for the Five Scoring Procedures	310
5.39 Situational Judgment Test: Summary of Item Analysis Results	311
5.40 Situational Judgment Test: Scores for Demo- graphic Subgroups	312
5.41 Situational Judgment Test: Combat/Non-Combat and MOS Differences in Scores	313
5.42 Situational Judgment Test: Interrater Reliabil- ities for Response Alternative Dimensional Ratings and Mean Rating and Mean Effectiveness for Each Dimension	315
5.43 Correspondence of Situational Judgment Test Dimensions With Job Analysis Dimensions	316
5.44 Situational Judgment Test: Computation of the Mean Effectiveness of Each Dimension	317
5.45 Situational Judgment Test: Mean Scores for Soldiers With Different Levels of Supervisory Training and Experience	318
5.46 Correlations Between Situational Judgment Test Scores and Supervisory Experience	319
5.47 Descriptive Statistics for Simulation Exercises . . .	323
5.48 Personal Counseling Exercise Items and Factor Analysis Results	325
5.49 Disciplinary Counseling Exercise Items and Factor Analysis Results	327
5.50 Training Exercise Items	328

CONTENTS (Continued)

	Page
Table 5.51 Confirmatory Factor Analysis Results for Simulation Exercises	329
5.52 Intercorrelations Between Scale Scores for Simulation Exercises	329
5.53 Army Job Satisfaction Questionnaire (AJSQ) Sample Description	331
5.54 AJSQ Principal Components Analysis	332
5.55 AJSQ Subscore Intercorrelation Matrix	333
5.56 AJSQ Means and Standard Deviations by Race and Gender	334
6.1 Correlations Among the CVII Summary Mea- sures Based on All Soldiers With Complete Data After Minimal Imputation	339
6.2 LISREL Results for Training and Counseling Factor Model: Parameter Estimates	342
6.3 LISREL Results for Training and Counseling Factor Model: Fit Statistics	343
6.4 LISREL Results for Training and Counseling Factor Model: t -Values	344
6.5 LISREL Results for Training and Counseling Factor Model: Fitted and Normalized Residuals	345
6.6 LISREL Results for "Overfit Model": Parameter Estimates	347
6.7 LISREL Results for "Overfit Model": Fit Statistics	349
6.8 LISREL Results for "Overfit Model": t -Values	350
6.9 LISREL Results for "Overfit Model": Fitted and Normalized Residuals	351
6.10 Means of the CVII Summary Measures for All Soldiers and for Soldiers With and Without Supervisory Experience	353

CONTENTS (Continued)

	Page
Table 6.11 Correlations Among the CVII Summary Measures Based on Soldiers With Supervisory Experience	354
6.12 Correlations Among the CVII Summary Mea- sures Based on Soldiers Without Supervisory Experience	355
6.13 Correlation of Specific Measures With Provisional Performance Scores for CVII	357

LIST OF FIGURES

Figure 1.1 Initial Project A organization	12
1.2 Initial Project A Governance Advisory Group	14
1.3 Project A research flow	16
1.4 Flow chart of predictor measure development activities of Project A	19
1.5 Building the Career Force: Initial project management structure	45
1.6 Glossary of terms for Project A/Career Force research samples	47
3.1 Experimental Predictor Battery Tests and Relevant Constructs	75
3.2 Summary of computer-administered test parameters . . .	105
3.3 Perceptual Speed and Accuracy Test decision time means by parameters	114
3.4 Comparison of Concurrent Validation and Longitudinal Validation composites	136
3.5 Comparison of ABLE composites for the Longitudinal and Concurrent Validations	159
3.6 Longitudinal Validation: Screening indexes developed for the AVOICE	165

CONTENTS (Continued)

	Page
Figure 3.7 Comparison of AVOICE composites for the Longitudinal and Concurrent Validations	176
3.8 Longitudinal Validation: Model for for- mation of JOB composites	189
3.9 Longitudinal Validation Experimental Battery: Composite scores and constituent basic scores	195
4.1 End-of-Training Army-wide performance rating scales	219
5.1 Example of supervisory/leadership performance ratings	249
5.2 Batch A MOS first-/second-tour criterion administration schedule	256
5.3 Hierarchical relationships among functional categories, task factors, and task constructs	290
5.4 Example rating items from Personal Counseling Simulation	321
5.5 Summary list of second-tour basic criterion scores	335
6.1 Relationship of specific variable to overall factors in the CVII performance model	356

BUILDING AND RETAINING THE CAREER FORCE: NEW PROCEDURES FOR ACCESSING AND ASSIGNING ARMY ENLISTED PERSONNEL. ANNUAL REPORT, 1990 FISCAL YEAR

Chapter 1 Introduction

This report is a summary of the major activities undertaken during the first 15 months of a Department of the Army research project entitled Building the Career Force. The report covers the period from 17 July 1989 through 30 September 1990. The research was conducted by a consortium composed of Human Resources Research Organization (HumRRO), American Institutes for Research (AIR), and Personnel Decisions Research Institute (PDRI), and the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI).

The research effort is the second phase of a two-phase program to develop a selection and classification system for enlisted personnel, based on expected future performance. Phase One was Project A. Its goals were to validate the Armed Services Vocational Aptitude Battery (ASVAB) by collecting data from a representative sample of Military Occupational Specialties (MOS), and to build a large and versatile data base by developing and validating new predictors and criterion measures that represented the entire domain of potential measures. The goals of Building the Career Force are to determine the longitudinal relationship between the new predictors and first-tour performance, to finalize and administer the measures of second-tour job performance, and to examine how selection and classification tests administered before a soldier's first enlistment, in conjunction with performance during that soldier's first enlistment, predict performance in a second enlistment.

The remainder of this chapter provides the military context for the two-phase research program, reviews the organization, objectives, and results of Project A, and describes the objectives and organization of Building the Career Force.

CHARACTERISTICS OF PRESENT ARMY PERSONNEL SYSTEM

The size, diversity, and widespread geographical distribution of Army activities have long dictated that the initial stages of personnel recruitment, selection, classification, and training be performed across many specialized units or activities and by personnel who have been specifically trained for these functions with guidance from command. Certain other functions are both formalized and carried out at the command level. These include unit or on-the-job training; performance evaluation; and decisions (or recommendations) concerning promotion, discipline, reassignment, and retention or separation from service. The major stages of the selection, classification, and assignment process for persons entering enlisted service in the Army are presented in Table 1.1. The stages are discussed below.

Recruitment

Recruitment, selection, and classification can hardly be discussed separately; they are interdependent processes. Their complementary nature should be evident in the ensuing description. The Army has succeeded in meeting or approximating its numerical recruitment quotas in most of the years

Table 1.1

The Army Selection, Classification, and Evaluation Process

<u>Stage/Activity</u>	<u>Processes^a</u>	<u>Outcomes^a</u>
Recruitment (U.S. Army Recruiting Command)	<ul style="list-style-type: none"> ● Recruiting Incentives, Options ● Recruiter Interviews ● Aptitude Prescreening Test (EST) (CAST) ● Records Checks 	<ul style="list-style-type: none"> ● To MET Sites or MEPS ● Disqualified
Center Selection/ Classification (Military Entrance Processing Station, MEPS)	<ul style="list-style-type: none"> ● Aptitude Testing (ASVAB) ● Physical Exam ● Moral Screening ● Special Tests ● Skill/Training Counseling ● Classification 	<ul style="list-style-type: none"> ● To Training ● Disqualified
Entry Training (Army Training Centers and Schools)	<ul style="list-style-type: none"> ● Basic Training ● Individual Training ● Training Evaluation ● Assignment ● Disciplinary Reviews ● Special Courses 	<ul style="list-style-type: none"> ● To Units ● Reassigned/Recycled ● Discharged
First Term (Operating Units)	<ul style="list-style-type: none"> ● Unit (on-the-job) Training and Mission Activities ● Special Courses ● Evaluation Tests, Ratings, Disciplinary Reviews ● Promotion Eligibility ● Reenlistment Counseling and Screening ● Army Continuing Education System 	<ul style="list-style-type: none"> ● Promotion/Demotion ● Discharged (prior to ETS) ● Separation (ETS) ● Reenlistment
Second Term (Operating Units)	<ul style="list-style-type: none"> ● Unit Training and Mission Activity ● Advanced Technical/Leadership Training ● Evaluation ● Promotion Eligibility 	<ul style="list-style-type: none"> ● Promotion/Demotion ● Reassigned ● Discharged (prior to ETS) ● Separation
(ETS)		<ul style="list-style-type: none"> ● Reenlistment

^a ASVAB = Armed Services Vocational Aptitude Battery; CAST = Computerized Adaptive Screening Test; EST = Enlisted Screening Test; ETS = Estimated Time of Separation; MET = Mobile Examining Team; SQT = Skill Qualification Test.

following the change to an All-Volunteer Force. This has resulted in an annual average of from 90,000 to 125,000 enlisted accessions from over twice as many applicants in the past 10 fiscal years. Furthermore, many qualified applicants do not begin active duty immediately but enter the delayed entry program (DEP) where they await a training slot.

The Army seeks to recruit the most capable personnel. Quality is generally defined in terms of high school graduation status and average or above scores on the Armed Forces Qualification Test (AFQT). The AFQT is a composite of four subtests (comprising verbal and math content) from the overall selection and classification instrument, the ASVAB. AFQT scores are reported in percentiles relative to the national youth population and grouped for convenience into the following categories and subcategories:

<u>AFQT Category</u>	<u>Percentile Score Range</u>
I	93 - 100
II	65 - 92
IIIA	50 - 64
IIIB	31 - 49
IVA	21 - 30
IVB	16 - 20
IVC	10 - 15
V	1 - 9

Categories I and II signify well above and above average trainability, respectively. Category III denotes average trainability, and Category IV signifies below average. Individuals scoring within Category V are, by law, ineligible for enlistment. Because of likelihood of success in training, the Army attempts to maximize the recruitment of those scoring within Categories I through IIIA. In addition, because traditional high school graduates are more likely to complete their contracted enlistment terms, in contrast to nongraduates and alternative credential holders (e.g., General Education Development (GED)), they are most actively recruited.

Though qualification for initial enlistment into the Army is based upon a number of criteria (including age, moral standards, and physical standards), education and particularly aptitude are the most pervasive and scrutinized criteria. The Army tries to target its advertising and aim its production recruiting resources to attract quality recruits. Also, as a means of identifying recruitment prospects while offering a career guidance tool, the ASVAB is administered to 900,000 high school juniors and seniors annually as part of the Department of Defense Student Testing Program.

The Army has recruited some non-high school graduates and applicants scoring in AFQT Category IV in order to meet numerical requirements and budget constraints. And, between 1976 and 1980, the Army erroneously enlisted high proportions of these less-preferred recruits as a result of a misnorming of the ASVAB (Maier and Truss, 1983). This situation raised concerns in Congress, and led to the imposition of ceilings on the proportion of non-high school graduates and Category IVs who may be enlisted. One of the outcomes of both Project A and Building the Career Force will be a much more solid empirical basis for qualification decisions.

To compete with the other Services and with the private sector for the prime target group, the Army has had to offer a variety of special inducements including "critical skill" bonuses and educational incentives. One of the most popular inducements has been the "training of choice" enlistment to a specific school training program, provided that applicants meet the minimum aptitude and educational standards and other prerequisites and that training "slots" are available at the time of their scheduled entry into the program. Additional options, offered separately or in combination with "training of choice," include guaranteed initial assignment to particular commands, units, or bases, primarily in the combat arms or in units requiring highly technical skills. In recent years, a large proportion of all Army recruits, particularly in the preferred aptitude and educational categories, have been enlisted under one or more of these options. An important research contribution would be to provide counselors with improved data-based aids to help create optimal person-job choices in light of Army manpower needs.

The importance of aptitude in recruiting decisions is exemplified in the prescreening of applicants at the recruiter level. For applicants who have not previously taken the ASVAB and whose educational/aptitude qualifications appear to be marginal based on the Army's trainability standards, the recruiter may administer a short Computerized Adaptive Screening Test (CAST) or Enlisted Screening Test (EST) to assess the applicant's prospects of passing the ASVAB. Applicants who appear upon initial recruiter screening to have a reasonable chance of qualifying for service are referred either to one of 1000 Mobile Examining Team (MET) sites for administration of the ASVAB, or directly to a Military Entrance Processing Station (MEPS) where all aspects of enlistment testing are conducted.

Selection and Classification at the Processing Station

Based on the information assembled, MEPS personnel complete classification and assignment to a particular training activity for applicants found qualified for enlistment.

The current versions of the ASVAB (Forms 15-17) consist of the following 10 subtests:

- Arithmetic Reasoning (AR)
- Numerical Operations (NO)
- Paragraph Comprehension (PC)
- Word Knowledge (WK)
- Coding Speed (CS)
- General Science (GS)
- Mathematics Knowledge (MK)
- Electronics Information (EI)
- Mechanical Comprehension (MC)
- Automotive-Shop Information (AS)

In addition to AFQT scores, subtest scores are combined to form 10 aptitude composite scores, based on those combinations of subtests that have been found to be most valid as predictors of successful completion of the various Army school training programs. For example, the composite score for electronics specialties is based on a combination of the scores for Arithmetic Reasoning, General Science, Mathematics Knowledge, and Electronics Information.

As stated above, eligibility for enlistment, in terms of the "trainability" standard, is based upon a combination of criteria: AFQT score, Aptitude Area composite scores, and whether the applicant is a high school diploma graduate. Under the most recent Army regulations,¹ the following standards are in effect:

- High school graduates are eligible if they achieve an AFQT percentile score of 16 or higher and a standard score of 85 in at least one Aptitude Area.
- GED high school equivalency holders are eligible if they achieve an AFQT percentile score of 31 or higher and a standard score of 85 in at least one Aptitude Area.
- Non-high school graduates are eligible only if they achieve an AFQT percentile score of 31 or higher and standard scores of 85 in at least two Aptitude Areas.

In addition to these formal minimum requirements, the Army may set higher operational cut scores for one or all of these groups.

Physical standards are captured in the PULHES profile, which rates the applicant on General Physical (P), Upper Torso (U), Lower Torso (L), Hearing (H), Eyes (E), and Psychiatric (S). Scores of 1 or 2 (on a 4-point scale) are required on all six indicators to be accepted for military duty (though waivers may be extended to applicants with a score of 3 on one or two indicators). The Army also sets general height and weight standards for enlistment.

Initial Classification

The overwhelming majority of Army enlistees enter the Army under a specific enlistment option that guarantees choice of initial school training, career field assignment, unit assignment, or geographical area. For these applicants, the initial classification and training assignment decision must be made before they enter service. This is accomplished at MEPS by referring applicants who have passed the basic screening criteria (aptitude, physical, moral) to an Army guidance counselor, whose responsibility is to match the applicant's qualifications and preferences to Army current skill training requirements and to make "reservations" for training assignments, consistent with the applicant's enlistment option.

For the enlistee, this decision will determine the nature of his or her initial training and occupational assignment, future military work environment, and chances of successful advancement in an Army career. For the Army, the relative success of the assignment process will significantly determine the aggregate levels of performance and attrition for the entire force.

The classification and training "reservation" procedure is handled by the Recruit Quota System (REQUEST), which was implemented in 1973. REQUEST is

¹Army Regulation 601-210, Regular Army and Army Reserve Enlistment Program, 1 October 1980, revised, Table 2-2.

a computer-based system designed to coordinate the information needed to reserve training slots for volunteers. REQUEST uses minimum qualifications for accessions control. Thus, to the extent that an applicant may minimally qualify for a wide range of courses or specialties, based on aptitude test scores, the initial classification decision is governed by (a) his or her own stated preference (often based upon limited knowledge about the actual job content and working conditions of the various military occupations), (b) the availability of training slots, and (c) the current priority assigned to filling each military occupational speciality (MOS).

These interactions among recruitment, selection, and classification in the current Army system give rise to several issues. There is an evident need for decision-making algorithms designed to maximize the overall utility of the MOS assignments. This requires that the average differential utilities of alternative assignments be known, as well as the marginal utility of each additional assignment to an MOS. The Army system currently incorporates marginal utilities by specifying desired distributions of AFQT scores, which are termed quality goals. In general, the parameters of recruit supply and demand (e.g., number of applicants in various categories, selection ratio, percentage of training slots filled, MOS priority) must also be taken into account when developing decision-making algorithms for selection and classification. The decision process must also allow for the potentially adverse impacts on recruitment if the enlistee's interests, work values, and preferences are not given sufficient weight. There are clear trade-offs that must be evaluated between the procedures necessary (a) to attract qualified people, and (b) to put them into the right slots.

Initial Training

After three days of processing at a Reception Battalion, all non-prior service Army recruits are assigned to a basic training (BT) program of 8 weeks. This is followed, with few exceptions, by a period of advanced individual training (AIT), designed to provide basic entry-level skills. Entrants into the combat arms and the military police receive both their basic training and their AIT at the same Army base (One Station Unit Training, OSUT) in courses of about 3-4 months' total duration. Those assigned to other specialties are sent to separate Army technical schools whose course lengths vary considerably, depending upon the technical complexity of the MOS. The diversity of course offerings is illustrated by the fact that the Army provides initial skills training in about 240 separate courses.²

In contrast to earlier practice, most enlisted trainees do not currently receive school grades upon completion of their courses, but are evaluated under Pass/Fail criteria. Those initially failing certain portions of a course are recycled. The premise is that slower learners, given sufficient time and effort under self-paced programs, can normally be trained to a satisfactory level of competence, and that this additional training investment is cost-effective. Those who continue to fail the course may be reassigned to other, often less demanding specialties or discharged from service.

²Department of Defense, Military Manpower Training Report for 1982, March 1981, p. II-4.

Performance Assessment in Army Units

Upon assignment to an Army unit, most of the personnel actions affecting the career of the first-term enlistee are initiated by his or her immediate supervisor and/or the unit commander. These include the nature of the duty assignment, the provision of on-the-job or unit training, and assessments of performance, both on and off the job. These assessments influence such decisions as promotion, future assignment, and eligibility for reenlistment, as well as possible disciplinary action (including early discharges from service).

To assure that these processes are administered fairly and consistently, in a manner compatible with broader Army objectives, the various aspects of enlisted personnel management are governed by detailed Army regulations. Army Regulation 600-211, The Enlisted Personnel Management System and related regulations cover such subjects as enlisted personnel evaluation and promotion, while AR 601-280, The Army Reenlistment Program prescribes the qualifications for reenlistment.

During an initial 3-year enlistment term, the typical enlistee can expect to progress to pay grade E-4, although advancement to higher pay grades for specially qualified personnel is not precluded. Authority to promote qualified personnel up to grade E-4 is delegated to unit commanders; promotion to higher grades is numerically restricted and must be approved either by field grade commanders for grades E-5 and E-6 or by Headquarters, Department of the Army for grades E-7 through E-9. Promotion to E-2 is almost automatic after 6 months of service. Promotions to grades E-3 and E-4 normally require completion of certain minimum periods of service (12 and 24 months, respectively), but are subject to certain numerical strength limitations and specific commander approval. Unit commanders also have the authority to reduce assigned soldiers in pay grade, based on misconduct or inefficiency.

The Enlisted Evaluation System provides for an evaluation both of the soldier's proficiency in his or her MOS and of overall duty performance. The process includes a subjective evaluation based on supervisory performance appraisal and ratings conducted at the unit level under prescribed procedures.

Reenlistment Screening

The final stage of personnel processing of first-term enlisted personnel is screening for reenlistment eligibility. As described in AR 601-280, this process considers such criteria as disciplinary records; Aptitude Area scores (based on ASVAB or its predecessors); low job evaluation scores, when applicable; and slow grade progression "resulting from a pattern of marginal conduct and/or performance." Enlisted personnel who do not meet certain minimum standards under these criteria must be approved before they can qualify for reenlistment.

The cumulative reductions due to attrition, reenlistment screening, and non-reenlistment of eligible personnel have resulted in the progressive diminution of initial Army cohorts to about 20-30 percent of their original numbers by the time they enter the fourth year of enlisted service. Not all of the latter, moreover, are retained or wish to be retained in their original

specialties, since an offer of retraining is often an inducement for reenlistment.

Summary

Even this brief description of the current system illustrates the complexity of the Army's personnel decision-making requirements and the large number of parameters that must be taken into account. In addition, decisions must be made for a very large flow of individuals within a very short time frame. In this regard the Army faces a much more difficult personnel management task than virtually any other organization. More effective selection/classification/promotion strategies would pay large dividends.

A BRIEF HISTORY OF SELECTION AND CLASSIFICATION

Formal personnel selection and classification using standardized measures of individual differences actually began in 1115 B.C. with the system of competitive examinations that led to appointment to the bureaucracy of Imperial China (DuBois, 1964). It soon included the selection/classification of individuals for particular military specialties, as in the selection of spear throwers with standardized measures of long-distance visual acuity (e.g., identification of stars in the night sky).

Systematic attempts to deal with selection/classification issues have been a part of military management ever since. Military organizations are virtually unique in their need to make large numbers of complex personnel decisions in a short space of time. However, the centrality of criterion-related validation to a technology of selection and classification was not fully articulated until World War II. Research and development sponsored by the military has been the mainstay of growth in that technology from then to the present.

The contributions of military psychologists during World War II are well-known and well-documented. The early work of the Personnel Research Branch of The Adjutant General's Office was summarized in a series of articles in the Psychological Bulletin (Staff, PRB, AGO, 1943 a, b, c, d, e, and f). Later work was published in Technical Bulletins and in such journals as Psychometrika, Personnel Psychology, and Journal of Applied Psychology. The Aviation Psychology Program of the Army Air Forces (AAF) issued 19 volumes, with a summary of the overall program presented in Volume I (Flanagan, 1948). In the Navy, personnel research played a smaller and less centralized role, but here too useful work was done by the Bureau of Naval Personnel (Stuit, 1947).

Much new ground was broken. Important advances were made in developing and analyzing criterion measures; Thorndike's textbook based on his Army Air Force experience presented a state-of-the-art classification and analysis of potential criteria (Thorndike, 1949). Rating scales were improved. Forced-choice methods were developed by the Personnel Research Branch; checklists based on critical incidents were used in the AAF program. The sequential aspect of prediction was articulated and examined; tests "validated" against training measures (usually pass/fail) were checked against measures of success in combat (usually ratings or awards). At least one "pure" validity study was accomplished, when the Air Force sent 1,000 cadets into pilot training without

regard to their pilot stanine derived from the classification battery. This remains one of the few studies that could report validities without correcting for restriction of range. Historically, 1940 to 1946 was a period of concentrated development of selection and classification procedures, and the further accomplishments of the next several decades flowed directly from it.

In part, this continuity is attributable to the well-known fact that many of the psychologists who had worked in the military research establishments during the war became leaders in the civilian research community after the war. In part, it is attributable to the less widely recognized fact that the bulk of the work continued to be funded by military agencies. The Office of Naval Research, the Personnel Research Branch (and its successors), and the Air Force Human Resources Research (HRR) installations were the principal sponsors.

The bibliography is very long. Of special relevance to the present project is the pioneering work on differential prediction by Brogden (1946a, 1951) and Horst (1954, 1955); on utility conceptions of validity by Brogden (1946b) and Brogden and Taylor (1950); on the "structure of intellect" by Guilford (1957); on the establishment of critical job requirements by Flanagan and associates (Flanagan, 1954); and on the decision-theoretic formulations of selection and classification developed by Cronbach and Gleser (1957) for the Office of Naval Research. The last of these (Psychological Tests and Personnel Decisions) was hailed quite appropriately as a breakthrough--a "new look" in selection and classification. But Cronbach and Gleser were the first to acknowledge the relevance of the work of Brogden and Horst cited above. It was the culmination of a lengthy sequence of development.

Project A was carried out in the context of this impressive history, and it has become another milestone. It was by far the most comprehensive personnel research and development project ever attempted. It was unique in that a complete personnel system was examined at one time. The jobs (MOS) examined were sampled representatively from the complete population, new predictor measures were sampled systematically from the complete domain of potential information, and job performance was assessed as thoroughly as possible with multiple measures. Given this data base, and using state-of-the-art analytic techniques, the functioning of the complete selection/classification decision process can be modeled and actually evaluated under various goals or constraints. Project A was truly a landmark in personnel research. The basis for this judgment is provided below in a summary description of Project A.

A SUMMARY DESCRIPTION OF PROJECT A

Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel, and Project B: An Enlisted Personnel Allocation System, were designed to provide the greatest possible increase in overall performance and readiness that can be obtained from improved selection, classification, and allocation of enlisted personnel. These two research programs provided an integrated examination of performance measurement, selection/classification, supply and demand parameters, and allocation procedures such that the Army could try to optimize the achievement of multiple personnel management goals (e.g., increase performance and decrease attrition).

The responsibilities of Project A were to develop: (a) a comprehensive set of new predictor measures; (b) multiple measures of job performance; (c) accurate estimates of the predictability of future performance; (d) decision rules for selection/classification at enlistment and reenlistment to optimize individual and system performance; and (e) a "what-if" gaming capability to illustrate the effects of variations in personnel management policies.

The point of departure for Project A was a decision by the Army in 1980 to initiate a long-term research effort to support the goal of an accessioning system that would base selection and classification decisions on expected future performance. The Deputy Chief of Staff for Operations, Department of Army (DCSOPS,DA) described the goals of the effort in a letter dated 19 November 1980.

The research effort initiated by DCSOPS, DA is being performed in two phases. Phase One was Project A. Its goal was to validate the ASVAB by collecting data from a representative sample of MOS and to build a large and versatile data base by developing and validating new predictors and criterion measures that represented the entire domain of potential measures. Phase Two is the Building the Career Force project. Its goals are, among other things, to determine the longitudinal relationship between new predictors and first-tour performance, to finalize and administer the measures of second-tour job performance, and to examine how well this performance is predicted by selection and classification tests administered before a soldier's first enlistment, in conjunction with measures of performance during that first enlistment.

Phase One (Project A) was an innovative and ambitious undertaking. Contrary perhaps to even the most optimistic expectations, it met virtually all its objectives and received high praise from its Scientific Advisory Group, its Army Advisory Group, and the National Academy of Sciences (NAS) Committee on the Performance of Military Personnel.

Project A Task Outline

Project A was designed as one integrated project organized into five major technical tasks, with a sixth task dealing with project management. The technical tasks were as follows:

Task 1 - Validation. Task 1 had two major components. The first component was to maintain the data base and provide the analytic procedures to determine the degree to which performance in Army jobs is predictable from some combination of new or existing measures. The second component was to conduct the appropriate analyses to determine whether the existing set of predictors, new predictors, or some combination of new and existing predictors has utility over and above the present system.

Task 2 - Developing Predictors of Job Performance. A large proportion of the efforts of the armed services in this regard have concentrated on improving the ASVAB, which is now a well-researched, valid measure of general cognitive abilities. However, many critical Army tasks appear to require psychomotor and perceptual skills for their successful performance. Further, neither biodata nor motivational variables were comprehensively evaluated. The objectives of Task 2 were to develop a broad array of new and improved

selection measures and to administer them to three major validation samples. A critical aspect of this task was to be the demonstration of the incremental validity added by new predictors.

Task 3 - Measurement of School/Training Success. The objective of Task 3 was to derive school and training performance indexes that could be used (a) as criteria against which to validate the initial predictors, and (b) as predictors of later job performance.

Task 4 - Assessment of Army-Wide Performance. In contrast to performance measures which may be developed for a specific Army MOS, Task 4 was to develop measures that could be used across all MOS (i.e., Army-wide). The intent was to develop measures of first- and second-tour job performance against which all Army enlisted personnel could be measured. A major objective was to develop a model of soldier effectiveness that specifies the major dimensions of an individual's contribution to the Army as an organization. Another important objective of Task 4 was to develop a procedure that could be used to scale the utility of performance.

Task 5 - Develop MOS-Specific Performance Measures. The focus of Task 5 was on the development of reliable and valid measures of specific job task performance for a selected set of MOS. This task consisted of three major components: job analysis, construction of job performance measures, and construct validation of the new measures. While only a subset of MOS were analyzed during this project, the Army may in the future wish to develop job performance measures for a larger number of MOS. For this reason, the methodology was to apply to all Army MOS.

The Organization of Project A

Initial Organization

The initial Project A organization is shown in Figure 1.1. The principal consortium task scientists are shown, with their respective organizations, in the lower row. The principal ARI scientists are shown in the upper row. Consortium and ARI scientists carried out research activities both independently and jointly. ARI scientists also had the administrative role of contract oversight. We include this diagram here only to show the matching of contractor and ARI staff and to illustrate the form of the project management and contract review structure. There were of course a number of personnel changes over the life of the project.

The Advisory Group Structure

A project of this scale had to maintain close and active coordination with the other military departments and the Department of Defense, as well as remain consistent with other ongoing research programs being conducted by the other Armed Services. The project also needed a mechanism for assuring that the research program met the highest standards for scientific quality. Finally, a method was needed to receive feedback from senior officers on priorities and objectives, as well as to identify current problems. An effective mechanism for meeting these needs was deemed to be a structure of advisory groups.

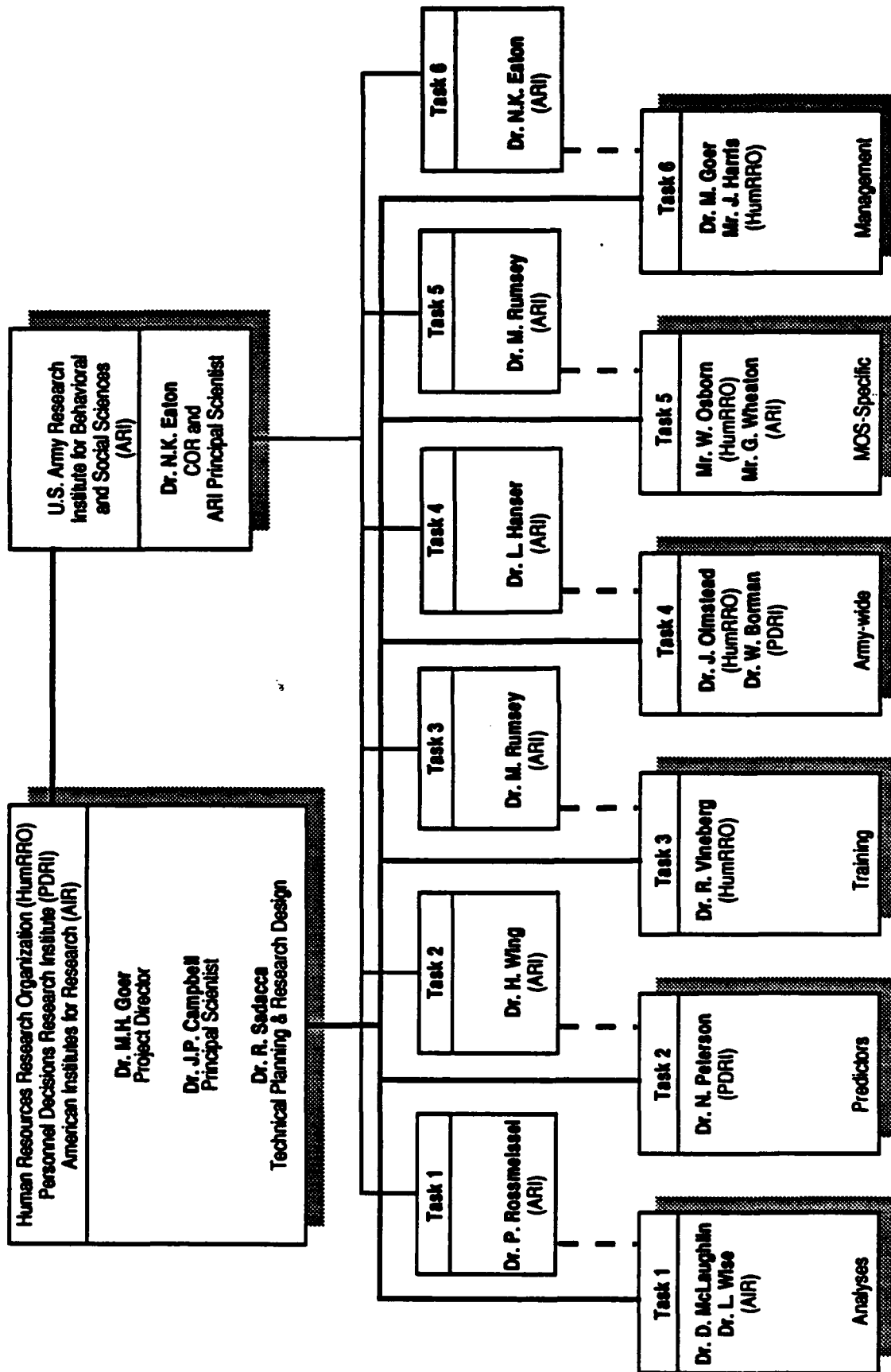


Figure 1.1. Initial Project A organization.

Figure 1.2 shows the structure and initial membership of the Governance Advisory Group (GAG) which comprised the Scientific Advisory Group (SAG), Inter-Service Advisory Group (ISAG), and General Officers Advisory Group (GOAG) components. The SAG was made up of nationally recognized authorities in psychometrics, experimental design, sampling theory, utility analysis, applied research in selection and classification, and the conduct of psychological research in the Army environment. It is perhaps indicative of the substance and success of Project A that all members of the Scientific Advisory Group remained with the Project from its beginning to the end.

The ISAG was comprised of the Laboratory Directors for applied psychological research in the Army, Air Force, and Navy, and the Director of Accession Policy from the Office of the Assistant Secretary of Defense for Manpower and Reserve Affairs. The GOAG included representatives from the Office of Deputy Chief of Staff for Personnel (DCSPER), Office of Deputy Chief of Staff for Operations (DCSOPS), Training and Doctrine Command (TRADOC), Forces Command (FORSCOM), and U.S. Army Europe (USAREUR).

The Research Plan and Integrated Master Plan

The first 6 months of Project A were spent in the planning, documenting, reviewing, modifying, and redrafting of research plans, troop support, administrative support, and budgetary plans, as well as in execution of initial research efforts. Drafts of the plans were provided to the SAG and ISAG. The culminating review was conducted in April 1983 by the General Officers Advisory Group, with representatives from the Scientific and Inter-Service Advisory Groups. The research program was endorsed by all three components of the GAG, and in May 1983, ARI issued Research Report 1332, Improving the Selection, Classification, and Utilization of Army Enlisted Personnel - Project A: Research Plan.

Specific Project A Objectives

The Project A Research Plan spoke to the specific operational and scientific outcomes that were to flow from the project.

Operational Objectives. The operational objectives were to:

- (1) Develop new measures of job performance that could be used as criteria against which to validate selection/classification measures.
- (2) Validate existing selection measures against both existing and project-developed criteria.
- (3) Develop and validate new selection and classification measures.
- (4) Develop a utility scale for different performance levels across MOS.

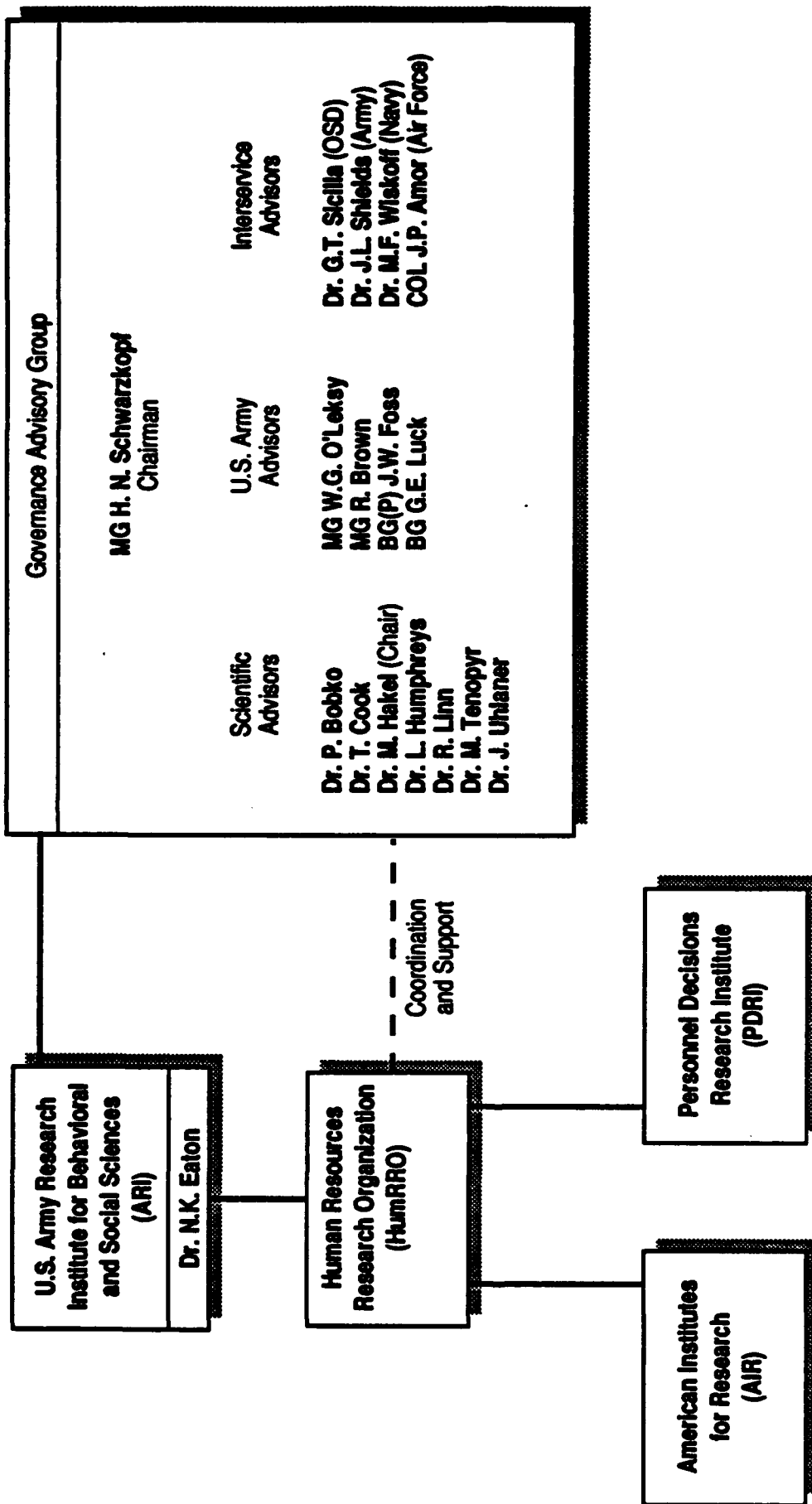


Figure 1.2. Initial Project A Governance Advisory Group.

Research Objectives. The research objectives were to:

- (1) Identify the constructs that constitute the universe of information available for selection/classification into entry-level skilled jobs.
- (2) Develop a general model of performance for entry-level skilled jobs.
- (3) Investigate the construct validity of the "method" variance in job performance measures.
- (4) Describe the utility functions and the utility metrics that individuals actually use when estimating "utility of performance."
- (5) Estimate the degree of differential prediction across (a) major domains of predictor information (e.g., abilities, personality, interests), (b) major factors of job performance, and (c) different types of jobs.

Research Design

The overall design of Project A used two predictive and one concurrent validation on two major troop cohorts (1983/1984 accessions and 1986/1987 accessions), and one file data validation on the 1981/1982 cohort. That is, in addition to collecting data from new samples, the project made use of existing file data for 1981 and 1982 accessions. Data from the accessions and Enlisted Master Files (EMF) were edited and merged into the Longitudinal Research Data Base (LRDB). A schematic of the data collection plan is shown in Figure 1.3.

The logic of the design was straightforward. Existing file data on the 81/82 cohort would provide an early opportunity to modify the existing operational selection and classification decision rules. In fact, the file data analyses were used to recommend changes in the composition of the ASVAB Aptitude Area composites.

The 83/84 cohort provided the first opportunity to obtain data using new predictor and performance measures. A "preliminary" battery of predominantly off-the-shelf tests provided new predictor data on soldiers in four MOS (05C [now 31C], 19E/K, 63B, 71L). These data, together with an exhaustive literature search, job analysis information, and multiple expert panel reviews, provided the information to construct a more tailored trial battery. This battery was administered concurrently with a variety of training, Army-wide, and MOS-specific performance measures in 1985 to the 1983/84 cohort.

The refinement of these measures resulted in the Experimental Predictor Battery, which was administered to a longitudinal sample from the FY86/87 cohort. The job performance criterion measures were administered to this cohort during late 1988. In addition, second-tour performance measures were developed for and administered to the FY83/84 cohort at the same time as part of a longitudinal followup of that sample into its second tour.

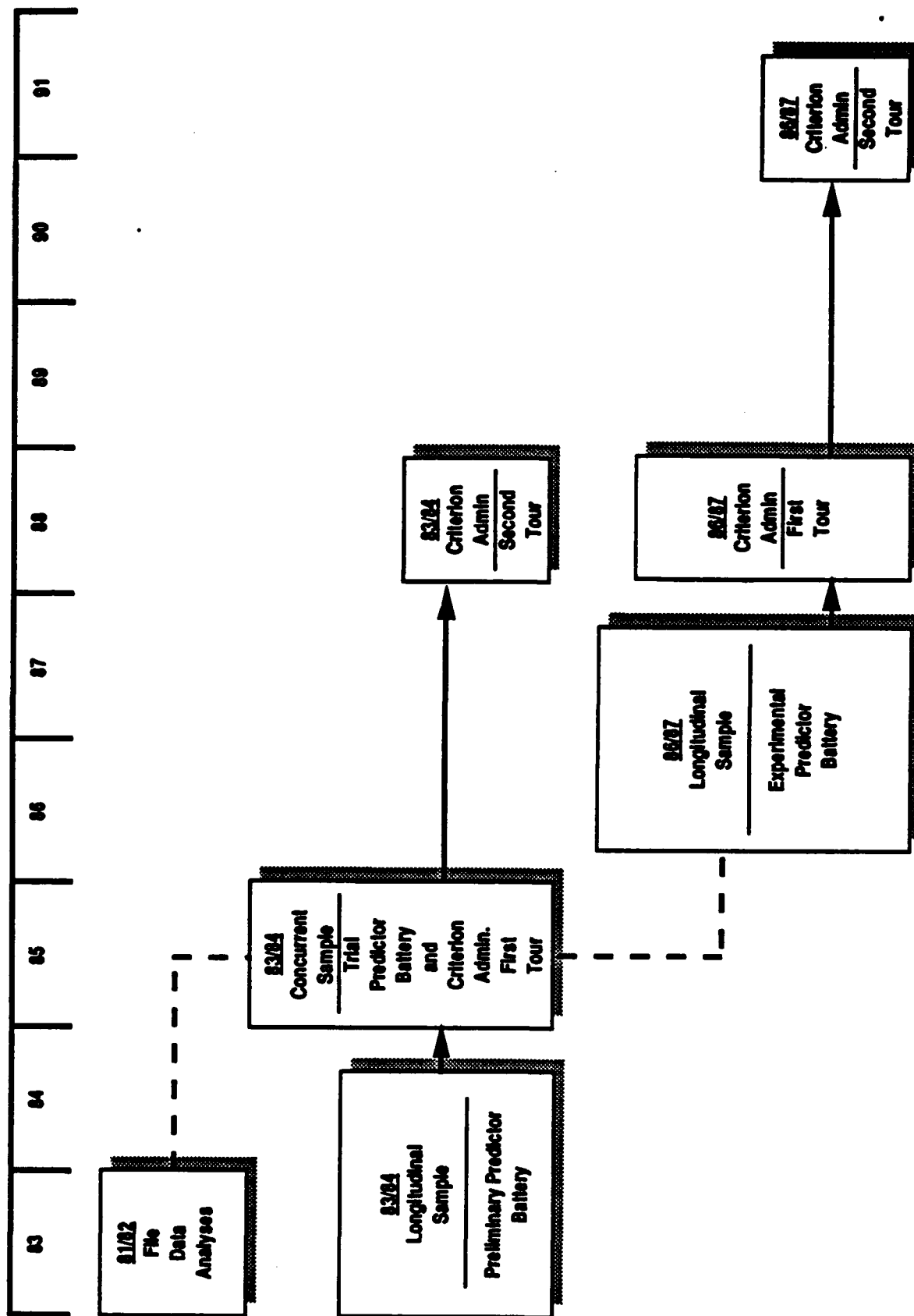


Figure 1.3. Project A research flow.

MOS and Sample Selection

The overall objective in generating the samples was to maximize the validity and reliability of the information to be gathered, while at the same time minimizing the time and costs involved. Costs are a function of the numbers of people in the sample, but are also influenced by the relative difficulty involved in locating and assembling the people in a particular sample.

The sampling plan itself incorporated two principal considerations. Because Project A was developing a system for a population of jobs (MOS), the MOS are the primary sampling units. However, there is a trade-off in the allocation of resources between the number of MOS researched and the number of subjects tested within each MOS: The more MOS are investigated, the fewer subjects per MOS can be tested, and vice versa. Cost versus statistical reliability considerations dictated that 19 MOS could be studied. The new predictors (from Task 2) along with the school and Army-wide performance measures (of Tasks 3 and 4) were administered to all 19. For nine MOS, the MOS-specific performance measures developed in Task 5 were also administered. These nine MOS were chosen to provide maximum coverage of the total array of knowledge, ability, and skill requirements of Army jobs, given certain statistical constraints.

The selection of the sample of 19 MOS proceeded through a series of stages. An initial sample of MOS was drawn by using the following considerations:

- (1) High-density MOS that would provide sufficient sample sizes for statistically reliable estimates of new predictor validity and differential validity across racial and gender groups.
- (2) Representative coverage of the Aptitude Areas measured by the ASVAB area composites.
- (3) High-priority MOS in the event of a national emergency, as rated by the Army.
- (4) Representation of the Army's designated Career Management Fields (CMF).
- (5) Representation of the jobs most crucial to accomplishment of the Army's mission.

On the basis of guidance from the Scientific Advisory Group, further refinements of the MOS sample were undertaken. These included a cluster analysis of expert ratings of MOS similarity and a review of the initial sample by the Governance Advisory Group. The similarity data were clustered and the initial results used to check the representativeness of the initial sample of 19 MOS. That is, did the initial sample of MOS include representatives from all the major clusters of MOS derived from the similarity scaling? On the basis of these results and guidance received from the Governance Advisory Group, two MOS that had been selected initially were replaced.

The sample of MOS resulting from the above procedures is shown in Table 1.2. The subsample of nine MOS to which the MOS-specific criterion measures were administered is shown as Batch A.

Table 1.2

Initial Project A Military Occupational Specialties (MOS)

<u>Batch A MOS</u>		<u>Batch Z MOS</u>	
05C ^a	Single Channel Radio Operator	12B	Combat Engineer
11B	Infantryman	16S	MANPADS Crewman
13B	Cannon Crewman	27E	TOW/Dragon Repairman
19E/19K ^b	Armor Crewman	51B	Carpentry/Masonry Specialist
63B	Vehicle & Generator Mechanic	54E	Chemical Operations Specialist
64C ^c	Motor Transport Operator	55B	Ammunition Specialist
71L	Administrative Specialist	67N	Utility Helicopter Repairer
91A	Medical Specialist	76W	Petroleum Supply Specialist
95B	Military Police	76Y	Unit Supply Specialist
		94B	Food Service Specialist

^aMOS 05C is now 31C.

^bMOS 19K was slated to replace 19E during the Longitudinal Validation. In practice, data were collected for both 19E and 19K throughout Project A.

^c64C is now 88M.

Predictor Development

A major objective was to develop an experimental battery of new selection/classification tests that would be potentially valuable additions to ASVAB and would maximize the Army's capability to make accurate selection/classification decisions. Consequently, the overall Project A strategy was to identify a universe of potential predictor constructs appropriate for the population of enlisted MOS, sample representatively from it, construct tests for each construct sampled, and refine and improve the measures through a series of pilot and field tests. The intent was to develop a predictor battery that was maximally useful for an entire population of jobs.

The long process of predictor development is represented in Figure 1.4. It began with an in-depth search of the entire personnel selection literature. Literature review teams were created for cognitive abilities, perceptual and psychomotor abilities, and non-cognitive characteristics such as personality, interest, and biographical history. Every available automated and manual technique was used in the search and an initial list of several hundred variables was compiled. The list went through several waves of expert review and was eventually reduced to a list of 53 potentially useful predictor variables. They are listed in Table 1.3.

A sample of 35 personnel selection experts was then asked to estimate the correlation between each predictor construct and each criterion factor, when that correlation was corrected for restriction of range and criterion unreliability. The resulting judgments could be analyzed for interjudge agreement, rows and columns could be factor analyzed, and the results could be compared to analogous information from the empirical literature. Most importantly, the exercise provided another substantial set of expert judgments

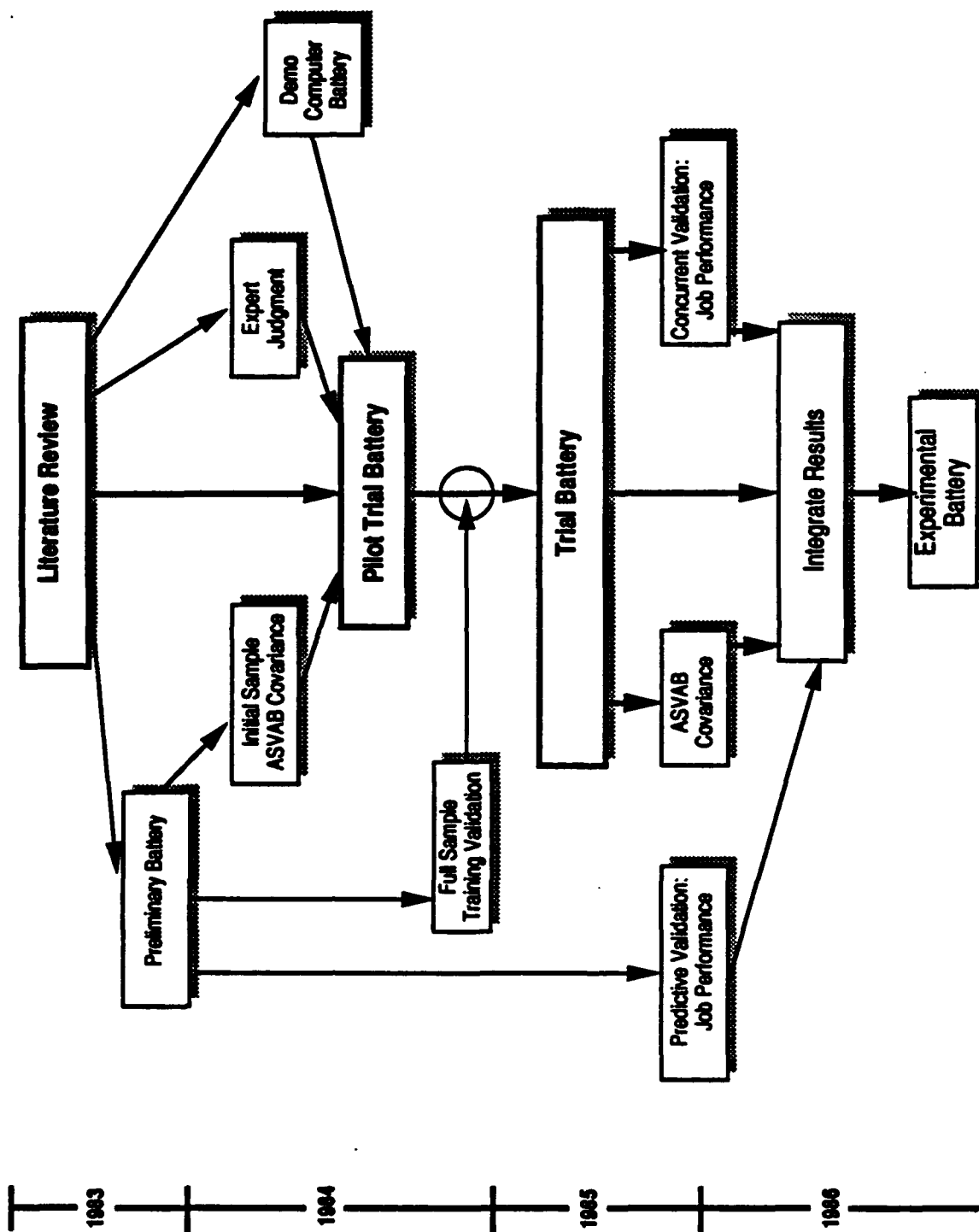


Figure 1.4. Flow chart of predictor measure development activities of Project A.

Table 1.3

Hierarchical Map of Predictor Space

Constructs	Clusters	Factors
1. Verbal Comprehension 5. Reading Comprehension 16. Identical Fluency 18. Analogical Reasoning 21. Omnibus Intelligence/Aptitude 22. Word Fluency	A. Verbal Ability General Intelligence	COGNITIVE ABILITIES
4. Word Problems 8. Inductive Reasoning Concept Formation	B. Reasoning	
2. Numerical Computation 3. Use of Formula/Number Problems	C. Number Ability	
12. Perceptual Speed and Accuracy	N. Perceptual Speed and Accuracy	
49. Investigative Interests	U. Investigative Interests	
14. Rote Memory 17. Follow Directions	J. Memory	
19. Figural Reasoning 23. Verbal and Figural Closure	F. Closure	
6. Two-dimensional Mental Rotation 7. Three-dimensional Mental Rotation 9. Spatial Visualization 11. Field Dependence (Negative) 15. Place Memory (Visual Memory) 20. Spatial Scanning	E. Visualization/Spatial	
24. Processing Efficiency 25. Selective Attention 26. Time Sharing	G. Mental Information Processing	
13. Mechanical Comprehension	L. Mechanical Comprehension	
48. Realistic Interests 51. Artistic Interests (Negative)	M. Realistic vs. Artistic Interests	MECHANICAL
28. Control Precision 29. Rate Control 32. Arm-hand Steadiness 34. Aiming	I. Steadiness/Precision	
27. Multilimb Coordination 35. Speed of Army Movement	D. Coordination	PSYCHOMOTOR
30. Manual Dexterity 31. Finger Dexterity 33. Wrist-finger Speed	K. Dexterity	
39. Sociability 52. Social Interests	Q. Sociability	SOCIAL SKILLS
50. Enterprising Interests	R. Enterprising Interests	
36. Involvement in Athletics and Physical Conditioning 37. Energy Level	T. Athletic Abilities/Energy	VIGOR
41. Dominance 42. Self-esteem	S. Dominance/Self-esteem	
40. Traditional Values 43. Conscientiousness 46. Non-delinquency 53. Conventional interests	M. Traditional Values/Convention- ability/Non-delinquency	MOTIVATION/ STABILITY
44. Locus of Control 47. Work Orientation	O. Work Orientation/Locus of Control	
38. Cooperativeness 45. Emotional Stability	P. Cooperation/Emotional Stability	

about which predictor constructs should be the most useful. A hierarchical analysis of the predictor validity profiles is also shown in Table 1.3.

All the available information was then used to arrive at a final set of variables for which new measures would be constructed. This represented months of effort by many people to select the variables that would best supplement the ASVAB in predicting job performance across all MOS. What followed were many months more of instrument construction, several waves of pilot tests, and a series of major field tests. Included in these efforts were the development of a computerized battery of perceptual/psychomotor tests, the creation of the software, the design and construction of a special response pedestal permitting a variety of responses (e.g., one-hand tracking, two-hand coordination), and the acquisition of 108 portable computerized testing stations. After each data collection, revisions were made on the basis of item statistics and expert review. Finally on 15 May 1985, the predictor battery was deemed ready for concurrent validation. That battery, known as the Trial Battery, is listed in Table 1.4.

Performance Measurement

The goals of measuring training performance and job performance in Project A were to define, or model, the total domain of performance in some reasonable way and then develop reliable and valid measures of each major factor.

Some additional specific goals were to (a) make a state-of-the-art attempt to develop job sample or "hands-on" measures of job task proficiency, (b) compare hands-on measurement to paper-and-pencil tests and rating measures of proficiency on the same tasks (i.e., a multitrait, multimethod approach), (c) develop standardized measures of training achievement for the purpose of determining the relationship between training performance and job performance, and (d) evaluate existing archival and administrative records as possible indicators of job performance.

Given these intentions, the criterion development effort employed three major methods: hands-on job sample tests, multiple-choice knowledge tests, and ratings. The behaviorally anchored rating scale (BARS) procedure was extensively used in developing the rating methods.

Modeling Performance

The development efforts were guided by a model that views performance as truly multidimensional. There is not one outcome, one factor, or one anything that can be pointed to and labeled as job performance. It is manifested by a wide variety of behaviors, or things people do, that are judged to be important for accomplishing the goals of the organization.

For the population of entry-level enlisted positions, two major types of job performance components were postulated. The first is composed of components that are specific to a particular job and that would reflect specific technical competence or specific job behaviors that are not required for other jobs. It was anticipated that there would be a relatively small number of distinguishable factors of technical performance that would be a function of different abilities or skills and would be reflected by different task

Table 1.4**Summary of Predictor Measures Used in Concurrent Validation (The Trial Battery)****COGNITIVE PAPER-AND-PENCIL TESTS**

<u>Test Name (Construct Name)</u>	<u>Number of Items</u>
Reasoning Test (Induction - Figural Reasoning)	30
Orientation Test (Spatial Orientation)	24
Map Test (Spatial Orientation)	20
Object Rotation Test (Spatial Visualization - Rotation)	90
Assembling Objects Test (Spatial Visualization - Rotation)	32
Maze Test (Spatial Visualization - Scanning)	24

COMPUTER-ADMINISTERED TESTS

<u>Name (Dimension)</u>	<u>Number of Items</u>
Simple Reaction Time (Processing efficiency)	15
Choice Reaction Time (Processing efficiency)	30
Memory Test (Short-term memory)	36
Target Tracking Test #1 (Psychomotor precision)	18
Target Shoot Test (Psychomotor precision)	30
Perceptual Speed and Accuracy Test (Perceptual speed and accuracy)	36
Identification Test (Perceptual speed and accuracy)	36
Target Tracking Test #2 (Two-hand coordination)	18
Number Memory Test (Number operations)	28
Cannon Shoot Test (Movement judgment)	36

NON-COGNITIVE PAPER-AND-PENCIL INVENTORIES

<u>Inventory Name and Subscale Name</u>	<u>Number of Items</u>
Assessment of Background and Life Experiences (ABLE)	209
Adjustment	
Dependability	
Achievement	
Physical Condition	
Leadership	
Locus of Control	
Agreeableness/Likability	
Army Vocational Interest Career Examination (AVOICE)	176
Realistic Interests	
Conventional Interests	
Social Interests	
Enterprising Interests	
Artistic Interests	

content. The second type includes components that are defined and measured in the same way for every job. These are referred to as Army-wide performance factors and incorporate the basic notion that total performance is much more than task or technical proficiency. It might include such things as contributions to teamwork, continual self-development, support for the norms and customs of the organization, and perseverance in the face of adversity.

In sum, the working model of total performance with which Project A began viewed performance as multidimensional within the two broad categories of factors or constructs. The job analysis and criterion construction methods were designed to explicate the content of these factors via an exhaustive description of the total performance domain, several iterations of data collection, and the use of multiple methods for identifying basic performance factors.

Saying that performance is multidimensional does not preclude using just one index to make a specific personnel decision (e.g., select/not select, promote/not promote). It seems quite reasonable for the organization to scale the importance of each major performance factor relative to a particular personnel decision that must be made, and to combine the weighted factor scores into a composite that represents the total contribution or utility of an individual's performance, within the context of that decision. The determination of the specific combinational rules (e.g., simple sum, weighted sum, nonlinear combination) that best reflect what the organization is trying to accomplish was to be a matter for research.

Criterion Development

Actual criterion development proceeded from two basic types of information. First, all available task descriptions were used to generate a population of job tasks for each MOS. The principal sources of task description are the Army Occupational Survey Program, which uses questionnaire checklists of several hundred task statements to survey job incumbents about the frequency with which they perform each task, and the Soldier's Manual for each job, which is a complete specification by management of what the task content of the job is supposed to be. After considerable editing, revision, and a formal review by a panel of subject matter experts, a population of 130-180 tasks was enumerated for each MOS in the sample.

An additional series of expert judgments was then used to scale the relative difficulty and importance of each task and to cluster tasks on the basis of content similarity. Sampling tasks for measurement was accomplished via a Delphi procedure. That is, each member of a team of task selectors was asked to select 30 tasks from the population of tasks such that the selected tasks were representative of task content, were important, and entailed a range of difficulty. The individual judge's choices were then regressed on the task characteristics and both the choices and the captured "policy" of each person were fed back to the group members, who then revised their choices as they saw fit. Typically, convergence was achieved quickly and the final selection was by consensus. The panel's selections were then thoroughly reviewed by the Army command responsible for that particular job.

Standardized hands-on job samples, paper-and-pencil job knowledge tests, and numerical ratings scales were then constructed to assess knowledge and

proficiency on these tasks. Each measure went through multiple rounds of pilot testing and revision.

The second procedure used to describe job content was the critical incident method. Panels of officers and noncommissioned officers (NCOs) generated thousands of critical incidents of effective and ineffective performance. There were two basic formats for the critical incident workshops. One asked participants to generate incidents that potentially could occur in any job. The second type focused on incidents that were specific to the content of the particular job under consideration. The behaviorally anchored rating scale procedure was used to construct rating scales for performance factors specific to a particular job (MOS-specific BARS) and performance factors that were defined in the same way and relevant for all jobs (Army-wide BARS).

The critical incident procedure was also used with workshops of combat veterans to develop rating scales of expected combat effectiveness.

Since a major project objective was to determine the relationships between training performance and job performance and their differential predictability, if any, a comprehensive training achievement test was constructed for each MOS. The content of the program of instruction (POI) was matched with the content of the population of job tasks, and items were written to represent each segment of the match. After pilot testing, revision, field testing, and Army proponent review, the result was a training achievement test of 150-200 items for each of the 19 MOS.

The final category of criterion measure was produced by a search of the Army's archival records for potential performance indicators. First, all possibilities were enumerated from three major sources of such records:

- The Enlisted Master File (EMF) - a central computer record of selected personnel actions.
- The Enlisted Military Personnel File (EMPF) - the permanent historical record of an individual's military service kept on microfiche at a central location.
- Military Personnel Records Jacket (MPRJ) - more commonly known as the 201 File, the personnel folder that follows the individual.

These three sources were systematically compared, using a sample of 750 people and a standardized information recording form. The 201 File looked the most promising in terms of recency and completeness, and six administrative performance indexes were eventually selected.

The complete array of performance measures, after revision on the basis of a large-scale field study of nine MOS (N = 150/MOS), is shown in Table 1.5. These are the measures which were administered to the concurrent sample of 400-600 people in each of the 19 MOS. The distinction between Batch A (9 MOS) and Batch Z (10 MOS) is that only Batch A MOS were given the job-specific tests; budget constraints dictated that not all criterion measures could be developed for each job and the job-specific measures were developed only for the nine MOS in Batch A.

Table 1.5

**Summary of Criterion Measures Used in Batch A and Batch Z
Concurrent Validation Samples^a**

Performance Measures Common to Batch A and Batch Z

- Army-wide rating scales (all obtained from both supervisors and peers).
 - Ten behaviorally anchored rating scales (BARS) designed to measure factors of non-job-specific performance
 - Single scale rating of overall effectiveness.
 - Single scale rating of NCO potential.
- Combat Performance Prediction scale (40 items).
- Paper-and-pencil Training Achievement Test developed for each of the 19 MOS (130-210 items each).
- Personnel File Information Form developed to gather objective archival records data (awards and letters, rifle marksmanship scores, physical training scores, etc.).

Performance Measures for Batch A Only

- Job-sample (hands-on) test of MOS-specific task proficiency.
 - Individual is test on each of 15 major job tasks in an MOS.
- Paper-and-pencil job knowledge tests designed to measure task-specific job knowledge.
 - Individual is scored on 150 to 200 multiple choice items representing 30 major job tasks. Ten to 15 of the tasks were also measured hands-on.
- Rating scale measures of specific task performance on the 15 tasks also measured with the knowledge tests. Most of the rated tasks were also included in the hands-on measures.
- MOS-specific behaviorally anchored ratings scales (BARS). From 6 to 12 BARS were developed for each MOS to represent the major factors that constituted job-specific technical and task proficiency.

Performance Measures for Batch Z Only

- Additional Army-wide rating scales (all obtained from both supervisors and peers).
 - Ratings of performance on 11 common tasks (e.g., basic first aid).
 - Single scale rating on performance of specific job duties.

Auxiliary Measures Included in Criterion Battery

- A Job History Questionnaire which asks for information about frequency and recency of performance of the MOS-specific tasks.
 - Army Work Environment Questionnaire - 53 items assessing situational/environmental characteristics, plus 46 items dealing with leadership.
 - Measurement Method Rating obtained from all participants at the end of the final testing sessions.
-

^a All rating measures were obtained from approximately two supervisors and three peers for each ratee.

The Concurrent Validation (CV)

Between 1 July and 1 December 1985, the predictor and criterion batteries were administered to 9,430 job incumbents in the 19 MOS. Four hours were devoted to the predictor tests and 12 hours to the criterion measures. Eight-person teams supported by four or five Army personnel visited each of 14 different Army posts for several weeks at a time. Considerable effort was devoted to training the data collection teams, standardizing testing conditions, keeping logs, and performing data checks each day. Table 1.6 shows the Concurrent Validation (CV) sample sizes by site and by MOS.

If all the rating scales are considered separately, the MOS-specific measures are aggregated at the task or instructional module level, and the major predictor subscales are used, approximately 200 criterion scores and 70 predictor scores were obtained on each individual. There was an obvious need to aggregate variables to reduce collinearity and make it easier to interpret the results appropriately.

For both predictors and criteria, the procedure for getting from the individual task or scale scores to factor or construct scores was similar except for the degree to which the previous literature was of help. Many decades of research on the measurement of abilities, personality, and interests have provided a lot of information about the structure of individual differences. Similar help from the performance side is really not available except for a modest number of descriptive studies of specific occupations such as managers, nurses, police officers, fire fighters, and college professors.

Given this initial disparity, both expert judgment and factor analytic results from the field tests were used to formulate hypothesized factors. These targets were then subjected to a series of quasi-confirmatory analyses using the Concurrent Validation sample. The resulting predictor construct scores and their associated component scales are shown in Table 1.7.

For the within-MOS criterion intercorrelation matrix, confirmatory analyses were used to test alternative models. The latent structure of performance that both fits the data in each job and seemed to make sense is portrayed in Table 1.8.

The model best confirmed by LISREL³ specified five substantive and two methods factors which were labeled the "ratings" factor and the "test" factor. The ratings factor was specified to be the first orthogonal component taken from all the ratings scales. The test factor is the first orthogonal component taken from the paper-and-pencil knowledge tests.

The first two substantive factors are based on the knowledge tests and the job sample measures. They are referred to as the Core Technical Performance factor and the General Soldiering Performance factor. The technical factor reflects content that is central and largely specific to the MOS. The general factor encompasses content that tends to be common across several jobs and is less central to the core performance objectives of each MOS.

³LISREL is a statistical software package that permits analysis of structural equation models (Joreskog & Sorbom, 1986).

Table 1.6

Concurrent Validation Sample Soldiers by MOS by Location

Location	Batch A MOS										Batch Z MOS										Total	Percent of Total
	11B	13B	19E	31C	63B	64C	71L	91A	95B	12B	16S	27E	51B	54E	55B	67M	76W	76Y	94B			
Fort Benning	45	23	41	7	13	39	16	9	13	13	15	3	0	12	18	9	13	15	12	316	3.35	
Fort Bliss	0	20	30	15	61	45	17	0	44	15	5	2	0	14	0	12	6	31	30	347	3.68	
Fort Bragg	68	46	0	0	37	25	41	10	72	82	75	13	19	72	20	7	42	39	62	730	7.74	
Fort Campbell	90	28	0	20	60	45	54	44	43	90	23	10	0	32	18	42	51	61	46	757	8.03	
Fort Carson	60	50	77	30	49	53	30	33	46	49	57	13	0	25	7	0	23	40	47	689	7.31	
Fort Hood	26	56	0	30	40	28	38	50	60	51	60	4	12	62	36	44	72	41	57	767	8.13	
Fort Knox	29	32	111	16	38	48	22	45	31	43	10	6	0	8	12	0	10	29	34	524	5.56	
Fort Lewis	75	46	13	11	43	46	23	27	56	27	25	1	11	51	31	20	48	41	36	631	6.69	
Fort Ord	30	0	0	14	30	42	31	43	51	51	7	8	1	4	7	15	23	40	28	425	4.51	
Fort Polk	73	47	19	29	47	47	18	46	44	60	45	9	8	16	7	23	26	51	35	648	6.87	
Fort Riley	30	43	55	27	26	45	35	30	40	31	20	8	8	25	52	0	20	39	45	579	6.14	
Fort Sill	0	108	0	20	43	51	44	0	29	42	11	0	0	0	0	15	7	35	32	437	4.63	
Fort Stewart	44	46	39	17	28	51	31	45	45	30	39	9	8	17	29	26	44	34	35	617	6.54	
USAREUR	132	122	120	130	122	121	114	119	118	120	78	61	41	96	54	63	105	134	113	1963	20.80	
Total	702	667	503	366	637	686	514	501	692	704	470	147	108	434	291	276	490	630	612	9430		
Percent Total	7.44	7.07	5.33	3.88	6.76	7.27	5.45	5.31	7.34	7.47	4.90	1.56	1.15	4.60	3.09	2.93	5.20	6.68	6.49			

Table 1.7

Predictor Construct Scores From Concurrent Validation Data

FROM COGNITIVE PAPER-AND-PENCIL TESTS

Overall Spatial Factor

Assembling Objects Test
Map Test
Maze Test
Object Rotation Test
Orientation Test
Figural Reasoning Test

FROM COMPUTERIZED MEASURES

Psychomotor Factor

Cannon Shoot Test (Time score)
Target Shoot Test (Time to fire)
Target Shoot Test (Log distance)
Target Tracking 1 (Log distance)
Target Tracking 2 (Log distance)
Short-Term Memory Test (Decision time)

Perceptual Speed/Accuracy Factor

Short-Term Memory Test (Percent correct)
Perceptual Speed & Accuracy Test (Percent correct)
Target Identification Test (Decision time)
Target Identification Test (Percent correct)

Number Speed/Accuracy Factor

Number Memory Test (Percent correct)
Number Memory Test (Initial decision time)
Number Memory Test (Mean operations decision time)
Number Memory Test (Final decision time)

General Reaction Speed Factor

Choice Reaction (Decision time)
Simple Reaction (Decision time)

General Reaction Accuracy Factor

Choice Reaction (Percent correct)
Simple Reaction (Percent correct)

FROM NON-COGNITIVE INVENTORIES

Achievement Factor

Self-Esteem scale
Work Orientation scale
Energy Level scale

Dependability Factor

Conscientiousness scale
Non-delinquency scale

Adjustment Factor

Emotional Stability scale

Physical Condition Factor

Physical Condition scale

Skilled Technician Interest Factor

Clerical/Administrative
Medical Services
Leadership/Guidance
Science/Chemical
Data Processing
Mathematics
Electronics Communications

Structural/Machines Interest Factor

Mechanics
Heavy Construction
Electronics
Vehicle/Equipment operator

Combat-Related Interest Factor

Combat
Rugged Individualism
Firearms Enthusiast

Audiovisual Arts Interest Factor

Drafting
Audiographics
Aesthetics

Food Service Interest Factor

Food Service - Professional
Food Service - Employee

Protective Services Interest Factor

Law Enforcement
Fire Protection

Table 1.8

Latent Structure Scores From Concurrent Validation Data

1. **Core Technical Proficiency: MOS (Job) specific core technical skills.** Indicates the proficiency with which the individual performs the tasks that are "central" to his or her job (MOS). The tasks represent the core of the job and they are its primary definers from job to job.
 - The subscales representing core content in both the knowledge tests and the job sample tests that loaded on this factor were summed within method, standardized, and then added together for a total factor score. The factor score does not include any rating measures.
 2. **General Soldiering Proficiency: General or common skills.** Covers a variety of general or common tasks -- e.g., use of basic weapons, first aid, which individuals in every MOS are responsible for being able to perform -- in addition to the core technical content specific to an MOS. This factor represents proficiency on these general tasks.
 - The same procedure (as for factor 1) was used to sum the general task scales, standardized within methods, and add the two standardized scores.
 3. **Peer Support and Leadership, Effort, and Self-Development.** Reflects the degree to which the individual exerts effort over the full range of job tasks, perseveres under adverse or dangerous conditions, and demonstrates leadership and support toward peers.
 - Five scales from the Army-wide BARS rating form (general technical performance, peer leadership, demonstrated effort, self-development, general maintenance), the expected combat performance scales, the job-specific BARS scales, and the total number of commendations and awards received by the individual were summed for this factor.
 4. **Personal Discipline.** Reflects the degree to which the individual adheres to Army regulations and traditions, exercises personal self-control, demonstrates responsibility in day-to-day behavior, and does not create disciplinary problems.
 - Scores on this factor are composed of three Army-wide BARS scales (adherence to traditions and regulations, exercising self-control, demonstrating integrity), a subscale from the combat rating pertaining to avoidance of trouble, and two indexes from the administrative records (disciplinary actions and promotions rate).
 5. **Physical Fitness and Military Bearing.** Represents the degree to which the individual maintains an appropriate military appearance and bearing and stays in good physical condition.
 - Factor scores are the sum of the physical fitness qualification score from the individual's personnel record and the "military bearing and appearance" rating scale.
-

The remaining factors are based on the ratings, primarily those developed by the critical incident method and the administrative/personnel records. Factor 3 encompassed the most scales and was the clearest in terms of its loading but appeared to be the most heterogeneous in terms of content. It seems to be a general effort and performance, performance under adverse conditions, peer leadership factor. Factor 4 is much more homogeneous and reflects the rating scales having to do with personal discipline and avoidance of trouble, and the number of negative personnel outcomes people reported. Factor 5 is fairly narrow in content and shows very clear loadings for ratings of military bearing and the physical fitness score that is part of everyone's personnel record. In general, this solution fit the data from all MOS and seemed reasonable and appropriate to Army management.

Concurrent Validation Results: Differential Prediction Across Criterion Components

The different criterion components were not predicted by the same things. Table 1.9 shows the multiple correlation of the components in these domains (corrected for shrinkage and for restriction of range, but not for unreliability) with the five criterion factors.

The entries in the table represent the average across all nine MOS. The level of validity of ASVAB for the first two factors is about the same as, or higher than, that usually observed when ASVAB is correlated with training criteria. ASVAB does predict job performance. For the third factor the validity of the cognitive tests drops, but is still substantial, and the validity of the non-cognitive inventories increases. This reversal becomes even more distinct for factors 4 and 5. Notice that the interest scales are also a reasonably good predictor of task performance and do not predict factors 3, 4, and 5 as well as the temperament scales.

Incremental Validity

An important question for the Army is how to improve on the validity of decisions made using the Army's current selection and classification instrument, the ASVAB. To help answer that question, the validity of the General Cognitive Ability scores (computed from the ASVAB) was compared to the validity obtained when the scores from a predictor domain were used to supplement the General Cognitive Ability composite. This was done for each performance construct within each of the nine jobs. Validities were then averaged across the nine jobs. The resulting mean incremental validities are reported in Table 1.10.

Relative Contribution of Individual Predictors

Because there were virtually no predictor by MOS interactions, a stepwise multiple regression solution within each of the six categories of predictor constructs was computed on the combined samples from the nine MOS in Batch A for each of the last four Army-wide performance factors (i.e., General Soldiering, Effort/Leadership, Personal Discipline, and Physical Fitness/Military Bearing).

Table 1.9

Mean Validity^a for the Composite Scores Within Each Predictor Domain Across Nine Army Enlisted Jobs

Job Performance Construct	Predictor Domain					Job Reward Preferences (K=3)
	General Cognitive Ability (K=4) ^b	Spatial Ability (K=1)	Perceptual- Psychomotor Ability (K=6)	Temperament (K=4)	Vocational Interests (K=6)	
Core Technical Proficiency	.63	.56	.53	.25	.35	.29
General Soldiering Proficiency	.65	.63	.57	.25	.34	.30
Effort and Leadership	.31	.25	.26	.33	.24	.19
Personal Discipline	.16	.12	.12	.32	.13	.11
Physical Fitness and Military Bearing	.20	.10	.11	.37	.12	.11

^aValidity coefficients were corrected for range restriction and adjusted for shrinkage.^bK is the number of predictor scores.

Table 1.10

Mean Incremental Validity^{a,b} for the Composite Scores Within Each Predictor Domain Across Mine Army Enlisted Jobs

	Predictor Domain					
	General					General
Job Performance Construct	General Cognitive Ability (K=4) ^c	General Cognitive Ability Plus Spatial Ability (K=5)	General Cognitive Ability Plus Perceptual- Psychomotor Ability (K=10)	General Cognitive Ability Plus Temperament (K=8)	General Cognitive Ability Plus Vocational Interests (K=10)	General Cognitive Ability Plus Job Reward Preferences (K=7)
	.63	.65	.64	.63	.64	.63
	.65	.68	.67	.66	.66	.66
	.31	.32	.32	.42	.35	.33
	.16	.17	.17	.35	.19	.19
Physical Fitness and Military Bearing	.20	.22	.22	.41	.24	.22

^aValidity coefficients were corrected for range restriction and adjusted for shrinkage.

^bIncremental validity refers to the increase in R afforded by the new predictors above and beyond the R for the Army's current predictor battery, the ASVAB.

^c K is the number of predictor scores.

Some comparisons of interest are the following:

- Among ASVAB scores the quantitative and technical scores contribute the most to the prediction of General Soldiering Proficiency. The verbal score plays a more prominent role in the prediction of the Core Technical performance factor.
- While ASVAB does not contribute much to the prediction of performance factors 4 and 5, the ASVAB technical score does make a relatively large contribution to the prediction of factor 3, the Effort/Leadership factor.
- The differential contributions of the temperament (ABLE) scores to prediction of performance factors 3, 4, and 5 are clear, significant, and pronounced. The profiles look like they should.
- The combat interests score was the most predictive interest score among the scores generated from the AVOICE.

The profile of regression coefficients for predicting the Core Technical Proficiency factor was significantly different across MOS, and the greatest differential was within the ASVAB and the AVOICE, and to a lesser extent within the spatial and computerized tests.

To look at the coefficients in another way, stepwise regressions were carried out when all 24 predictor scores were used to predict each performance factor. The analyses for the four Army-wide criterion factors were carried out on a combined sample while the analyses against the Core Technical factor were done MOS by MOS. Again the differential patterns appear across the four Army-wide performance factors and across MOS for the Core Technical factor. However, a surprise was the strong role played by the spatial factor and the combat interest factor in predicting the technical performance factor in the combat specialties.

Weighting Criterion Composites

The Concurrent Validation results indicated that each of the five criterion components can be predicted with considerable validity and that the validity of the different predictor domains varies systematically across criterion components. A subsequent focus was on the best method for obtaining their relative importance weights when the five components are combined into an overall composite index of performance. Consequently, weighting judgments were systematically gathered from carefully chosen samples of both NCOs and officers familiar with each MOS.

The five Project A performance constructs received significantly different patterns of weights in different MOS and the different groups of experts agreed, in general, on the relative ranking of the weights. For example, the Effort and Leadership construct tended to be rated highest among the combat MOS.

Multiple judges per MOS, about 30 on the average, produced average rater reliabilities that are quite respectable (above .95 for most MOS). High intermethod correlations (about .95 on the average) between the construct

weights obtained by a direct estimation method and a conjoint scaling method for the separate MOS further document the reliability of the means of the weights.

Scaling the Utility of Individual Performance

The utility problem for Project A was one of assigning utility values to MOS by performance level combinations. That is, if it is true that personnel assignments will differ in value to the Army, depending on the specific MOS to which an assignment is made and on the level at which an individual will perform in that MOS, then the value of a classification strategy that has a validity significantly greater than zero will increase to the extent that the differential values (utilities) can be estimated and made a part of the assignment system.

The general procedure used to obtain utility scale values for different levels of predicted performance in each MOS used field grade officers as expert judges and was divided into three phases. Phase one was exploratory and used a series of workshop meetings with various officer groups to uncover the major issues. The goal of phase two was to evaluate alternative expert judgment scaling methods and develop the procedure to be used. In phase three the selected methods were used to obtain the final scale values.

Perhaps the most significant finding was that Army officers would be willing and able to assign differential utility values across MOS and performance levels. Perhaps the next most significant finding was that stable scale values could be obtained from averaging across a relatively small number of officer judges.

The analyses supported the conclusions that (a) for both methods the reliability of the average value produced by 11 judges or more is very high; (b) reliabilities are high even when performance level is controlled and differences are due only to MOS differences within performance level; (c) judges from different posts or MOS backgrounds do not produce different patterns of scale values; and (d) within the limits of the methods used, the 1,365 MOS by performance level combinations have been placed on the same ratio scale of judged utility.

However, a number of problems need to be addressed before utilities similar to the ones obtained in Project A can be used operationally. One problem concerns the optimal distribution within MOS, considering both within- and between-MOS utilities as well as the available recruit pool and the quality of existing personnel. This is the issue of average vs. marginal utility (Nord & White, 1988). Another issue concerns the duration of time that the recruits actually remain in the Army and how to aggregate values over time.

The Longitudinal Validation (LV) Data Collections

The Longitudinal Validation began with the administration of the Experimental Predictor Battery at the reception battalions to more than 50,000 accessions from the 86/87 cohort. It then followed these soldiers through their Advanced Individual Training or One Station Unit Training, where they were administered several criterion measures of performance during training.

They were then followed into their first tour, where the job performance measures were administered.

In order to cover the ASVAB Aptitude Area composites more comprehensively, two MOS were added to the Batch Z domain: 29E, Communications Electronics Repairer and 96B, Intelligence Specialist. In addition, 19E/19K was split into two distinct MOS for measures development, and 76W was dropped because it was redundant with 76Y. These changes resulted in 21 MOS in Project A for the Longitudinal Validation.

The Experimental Predictor Battery

The predictor testing sites and the data collection period for each site were as follows:

<u>Site</u>	<u>Predictor Testing Period</u>
Fort Sill	20 Aug 86 - 20 Aug 87
Fort Benning	27 Aug 86 - 27 Aug 87
Fort Bliss	4 Sep 86 - 4 Sep 87
Fort Knox	10 Sep 86 - 10 Sep 87
Fort McClellan	17 Sep 86 - 17 Sep 87
Fort Dix	24 Sep 86 - 24 Sep 87
Fort Leonard Wood	1 Oct 86 - 1 Oct 87
Fort Jackson	19 Nov 86 - 19 Nov 87

Table 1.11 shows the complete array of tests and inventories in the Experimental Battery, the number of items in each, and the time limit (for the timed tests) or approximate time to finish (for the untimed inventories).

The information obtained from the Concurrent Validation (CV) data analysis was used to make the final revisions to the Predictor Battery for the LV. Since the battery had already been through several iterations of data collection, analysis, and revision, the revisions were not substantial.

Training Performance Measures

Measures of training performance were collected on each individual at the end of AIT. The measures consisted of a number of the Army-wide BARS scales collected from the individual's drill instructor and the training achievement test previously developed for each MOS.

Second-Tour Performance Criterion Development

Over the course of its life cycle, Project A was able to complete the necessary job analyses and begin the criterion development work for the assessment of second-tour NCO performance for the Batch A MOS.

The specific goals of the job-analytic work were to:

- Describe the major differences between entry-level and second-tour performance content, within MOS.

Table 1.11

Description of Tests in the Experimental Predictor Battery

	<u>Number of Items</u>	<u>Time Limit (minutes)</u>
COGNITIVE PAPER-AND-PENCIL TESTS		
Reasoning Test	30	12
Object Rotation Test	90	7.5
Orientation Test	24	10
Maze Test	24	5.5
Map Test	20	12
Assembling Objects Test	36	18
	<u>Number of Items</u>	<u>Approximate Time</u>
COMPUTER-ADMINISTERED TESTS		
Demographics	2	4
Reaction Time 1	15	2
Reaction Time 2	30	3
Memory Test	36	7
Target Tracking Test 1	18	8
Perceptual Speed and Accuracy Test	36	6
Target Tracking Test 2	18	7
Number Memory Test	28	10
Cannon Shoot Test	36	7
Target Identification Test	36	4
Target Shoot Test	30	5
	<u>Number of Items</u>	<u>Approximate Time</u>
NON-COGNITIVE PAPER-AND-PENCIL INVENTORIES		
Assessment of Background and Life Experiences (ABLE)	199	35
Army Vocational Interest Career Examination (AVOICE)	182	20
Job Orientation Blank (JOB)	31	5

- Describe the major differences across MOS, within the second-tour jobs.
- Describe the specific nature of the supervisory/leadership component of these higher level jobs.

Once these objectives were achieved, the information was used to address four questions:

- (1) What should be the content of the new criterion measures?
- (2) What kinds of measurement methods are needed?
- (3) Are separate measures needed for each job? Or are the jobs so similar that the same measures can be applied to all?
- (4) To what extent can measures developed for entry-level soldiers be used among higher level soldiers?

By Army policy⁴, all soldiers are responsible for being able to perform all tasks at lower skill levels, as well as the tasks at their skill level. Because of these policies, the first-tour job analyses were used as a starting point and additional job analysis information was collected to describe the second-tour changes. In addition the issue of leadership/ supervision performance was of special concern.

To capture both the technical and the supervisory aspects of an MOS, four methods of job analysis were used: task analysis, a standardized questionnaire measure of supervisory/leadership responsibilities, critical incident analyses, and interviews with small groups of senior NCOs.

Given available resources, constraints on testing time, guidance from the literature, previous Project A work, and the second-tour job analysis results, a potential set of measurement methods was identified and reviewed by the project staff and the Scientific Advisory Group. Some of the measurement methods had been used for the first tour and some were newly developed. As Project A was drawing to a conclusion, field test versions of the second-tour criterion measures were available.

Briefly, the array of second-tour measures included the following:

- (1) The original first-tour Army-wide behavioral and combat performance rating scales as modified on the basis of the second-tour job analysis.
- (2) The MOS-specific rating scales as modified by the second-tour job analysis.
- (3) Hands-on tests for 8-15 tasks for each MOS.
- (4) Job knowledge tests for approximately 30 tasks for each MOS.

⁴Army Regulation 611-201, Enlisted Career Management Fields and Military Occupational Specialties.

- (5) A self-report measure of administrative indexes, as modified by a reexamination of the records available for second-tour incumbents.
- (6) The Situational Judgment Test, a paper-and-pencil test designed to measure knowledge of what action to take in a series of critical supervisory/leadership situations.
- (7) Role Play Simulations
 - An exercise involving the counseling of a soldier with a personal problem.
 - An exercise involving the counseling of a soldier with a performance problem.
 - A simulation of one-on-one remedial training.
- (8) A set of rating scales designed to reflect the major dimensions of supervisory/leadership performance.

The above measures were administered to the 83/84 cohort second-tour followup sample in the fall and winter of 1988/89.

Job Performance Measurement

Project A concluded with the LV administration of the performance measures to both first-tour and second-tour incumbents between July 1988 and February 1989. The first-tour criterion measures are essentially the same as those used for the Concurrent Validation. The second-tour measures are the prototypes described in the preceding section.

Summary

At this point, Project A had reached its basic goals.

- Multiple criterion measures had been developed and used to formulate five components of job performance.
- ASVAB was shown to be a highly valid predictor of job performance as reflected in the Core Technical performance and General Soldiering performance components.
- There was considerable differential prediction for the total test battery across the five performance components within each MOS.
- The non-cognitive predictors added significantly to the prediction of the "will-do" components of performance and should prove to be valuable additions to the total system.
- As was expected, differential prediction across MOS was limited largely to the Core Technical performance factor. Both the ASVAB and the new experimental cognitive tests should contribute to differential prediction equations across major MOS clusters.

However, the full analyses necessary to determine the prediction equations remain to be done.

- The importance weights for the five performance components have been scaled within each of the 19 MOS.
- Reliable scale values have been obtained for comparing the average utility of 1,365 MOS (273) by performance level (5) combinations.
- Comprehensive job analyses and prototypic criterion measures have been completed for second-tour NCO performance in nine MOS.

The results are impressive; however, for their full benefit to be realized a number of things must happen. Both the covariance structures and the estimates of predictive validity must be cross-validated with a genuine predictive design (i.e., the Longitudinal Validation); the rules for forming criterion composites must be developed; the marginal utility of accurate predictions must be estimated; valid measures of NCO performance must be constructed; an NCO performance model must be developed; the specifics of the full selection/classification/promotion decision system must be modeled; and the effects of using the new predictors in various combinations under a variety of goals and constraints must be evaluated.

The results of the work performed during Project A are presented in a series of Annual Reports, as follows:

FY83, ARI Research Report 1347 and its Technical Appendix,
ARI Research Note 83-37;
FY84, ARI Research Report 1393 and two related reports,
ARI Technical Report 660 and ARI Research Note 85-14;
FY85, ARI Technical Report 746 and ARI Research Note 87-54;
FY86, ARI Technical Report 792 and ARI Research Note 88-36;
FY87, ARI Technical Report 862 and ARI Research Note 88-23;
FY88, ARI Research Note 91-34;
Final Report, ARI Research Report 1597.

BUILDING ON PROJECT A

Project A was designed according to certain specifications and followed a particular overall strategy. In general, the basic strategy was to build a large, sophisticated, and versatile research data base by developing new predictors and criterion measures that represented the entire domain of potential measures and by collecting validation data from a representative sample of MOS. The ingredients of this basic foundation are outlined below and the outcomes that resulted from Project A are summarized.

The Foundation Provided by Project A

The first three years of Project A were devoted largely to completing a comprehensive series of developmental steps. Much time, effort, and resources were devoted to predictor development, performance criterion development, and model development. It was indeed a very large initial investment in project research and development and the actual production of the major operational

products was deferred. Another year and one half was devoted to planning and conducting the first major data collection (the Concurrent Validation) and completing the basic data analysis. During the remainder of the project, data analysis continued, utility values for performance outcomes were estimated, criterion component weights were obtained, and the Longitudinal Validation data collection was initiated.

Although the time spent in the initial R&D phase created some delays in gratification, in retrospect it seems to have been a wise decision. This long development process can be thought of as a large initial investment which will in the end produce a much larger return than if the objectives had been pursued in more piecemeal fashion and the Army had tried for more short-run gains. As shown by the preceding summary of Project A, some of the payoff has been realized already. However, the major portion of the profit, particularly as it pertains to optimizing the entire selection/classification/promotion decision-making procedure, is yet to come. It is incumbent upon the Building the Career Force project to take maximum advantage of what Project A has accomplished. If this can be done effectively, the Army should realize very large gains from this initial investment.

For example, during the development phase of Project A, a 4-hour battery of new selection/classification tests was developed so as to sample systematically the most relevant applicant characteristics not presently covered by ASVAB. Also during the development phase, a 12-hour training achievement and job performance measurement procedure was constructed to provide multiple measures of every major component of performance for each job in a group of MOS representatively sampled from the population of entry-level MOS. Consequently, for (a) jobs, (b) performance components, and (c) selection/classification measures, a population had been defined and then sampled systematically.

This foundation makes the results of the Concurrent and Longitudinal Validations generalizable and extremely useful for guiding future selection/classification practices. A wide variety of comparative "what if" questions can be asked about the differential prediction (by different kinds of test information) of each major performance component under varying sets of constraints, and the answers generalized to the entire system. No other organization in the world, public or private, has such an extensive, systematically developed, and generalizable body of information with which to build and evaluate future selection, classification, and promotion strategies. It can be used for many years to come.

In addition to developing a comprehensive battery of selection/classification tests and a full multiple-method array of first-tour performance measures and using them to generate the most extensive data base in the history of personnel research, Project A yielded a number of both scientific and applied products. These are summarized below.

Project A Products and Results

The products in the following list are of two general kinds: products for the "science" (personnel research) and products for the organization (the Army). The list is intended to move from the scientific to the applied.

However, the distinction is not always easy to make since many products are useful for both.

- (1) There exist, in technical report form, comprehensive reviews of all validity evidence pertaining to selection and classification for skilled jobs. These are the most comprehensive such reviews ever done.
- (2) The question of whether ASVAB does or does not predict job performance (in addition to training performance) has been answered definitively, in the affirmative. The Army and the Department of Defense are now in a firmer position to support their quality goals. In addition, it is now known what aspects of performance ASVAB predicts best and which aspects of performance could be predicted better with other types of selection instruments.
- (3) A set of new experimental tests has been developed to measure non-cognitive, psychomotor, perceptual, and cognitive characteristics that are not now measured by the ASVAB. The scope of Project A made it possible to examine virtually the entire domain of selection information, sample from it, and investigate the basic incremental validity produced by each major piece of information.
- (4) Using much more comprehensive samples than ever before, new ASVAB Aptitude Area composites have been developed which are firmly data based and empirically defensible.
- (5) The results of an expert judgment study of expected correlations between predictor constructs and performance factors are available. In brief, a large sample of personnel experts considered the population of predictor and criterion variables appropriate for entry-level jobs and forecasted what the validity coefficients would be. The consistency in the judgments and their correspondence with known data points make these a potentially valuable tool for future test selection and synthetic validation work.
- (6) Much has been learned about the nature of performance in entry-level skilled jobs (e.g., first-tour MOS). We now have a much clearer idea of what major factors constitute performance and how they can be measured. The "criterion problem" is better understood. This knowledge base should better inform future enlistment and promotion policy, as well as future personnel research.
- (7) The Concurrent Validation data support the assertion that supervisor ratings of subordinate performance have considerable construct validity if a careful measurement procedure is followed. The data also support the conclusion that supervisors seem to assess both the technical performance of individuals and their general dependability/motivation at the same time.
- (8) Within the limits of the Concurrent Validation design, the incremental validity of appropriate ABLE scales for predicting the "will do" components of performance has been demonstrated.

- (9) The potential of the AVOICE for differentially predicting "can do" performance in combat vs. technical vs. administrative support MOS has been established. What is needed to make this finding operational is empirical scoring keys.
- (10) The Project A job/task analysis procedures worked well and can be used by the Army in the future to develop training curricula, performance measures, and field exercises. The job analysis summaries for each MOS serve as a model for future job analysis work in the Army as well as in the public and private sectors.
- (11) AIT training achievement measures have been developed for 21 MOS. The training measures will allow a determination of whether training performance predicts job performance, and whether it does so differentially for different groups of trainees (race, gender), and different groups of MOS (combat, combat support, combat service support).
- (12) The package of rating scale administration procedures can be used in future personnel research in the Army. A major effort in the Project A research was to develop an effective and very efficient set of procedures for administering performance rating scales to large numbers of people. These procedures and the package of materials can be adapted for use in other Army personnel research where ratings of many persons are required.
- (13) The Supervisory Description Questionnaire (which came out of second-tour job analyses work) is a very useful instrument for future work in the design of leadership training or the evaluation of leadership/supervisor performance. The questionnaire is based on a clear rationale and is straightforward to use.
- (14) Project A developed a common utility scale for making comparisons across MOS and performance levels within MOS. Although it does not speak to marginal utility issues, it can be used to enhance the comparison of alternative selection/classification procedures.
- (15) One very real, and very important product, is the Project A data base itself. It is by orders of magnitude the largest and most completely documented personnel research data base in existence.

BUILDING THE CAREER FORCE

The Project A data base, the predictor and criterion measures it developed, the working models it provided, and its basic analytic work have provided a valuable foundation for the further production of scientific findings and operational products, and for the subsequent investigation of reenlistment decisions, NCO job performance, NCO promotion decisions, and the identification of NCO potential.

In general, the work encompassed by this project to build and retain the career force is intended to accomplish several goals.

- (1) Build the final pieces required for a complete selection/classification decision-making system.
- (2) Provide the analytic procedures and data required to maximize the system's performance and evaluate its effectiveness.
- (3) Build the foundation for its implementation.

The principal focus is on the greatest possible gains in overall individual performance for both "can do" and "will do" components of performance that can be obtained from enhancing the selection/classification system for first- and second-tour enlisted personnel. Maximizing the benefit from a more effective match of people and jobs has always been a goal of the Army personnel system. Given the population demographics for the United States during the coming decade, this goal becomes even more crucial. It is incumbent on virtually every organization to go as far as the state-of-the-art will allow.

This means that the information that is used to make personnel decisions must yield the maximum gain in terms of accuracy and fairness of predictions. It means that the models and procedures used to execute selection and classification decisions must both serve the goals of the organization and maximize the aggregate benefits that can be obtained from using the selection/classification measures (e.g., new computerized tests). It means that the implementation of the system, or any part of it, must serve the needs of the users and also maintain fidelity with the goals on which the system is based.

Project Objectives

The specific objectives of the Career Force project are to:

- (1) Develop a complete array of valid and reliable measures of second-tour performance as an Army NCO, using the Project A prototypes as a starting point.
- (2) Carry out a complete incremental predictive validation of (a) the ASVAB and the Project A Experimental Battery of predictors, (b) measures of training success, and (c) the full array of first-tour performance criteria developed as part of Project A. The criteria against which these three sets of predictors will be validated, both individually and incrementally for each major criterion component, are the second-tour job performance measures.
- (3) Develop a "model" of second-tour NCO performance that parallels the first-tour performance model (from Project A) and that identifies the major components of second-tour performance, provides information on their construct validity, and specifies how the major components of performance should be combined for specific prediction or interpretation purposes.
- (4) Develop the analytic framework needed to evaluate the optimal prediction equations for predicting (a) training performance; (b) first-tour performance; (c) first-tour attrition and the

reenlistment decision; and (d) second-tour performance, under the conditions when testing time is limited to a specified amount and when there must be a trade-off among alternative selection/classification goals (e.g., maximizing aggregate performance vs. minimizing discipline and low-motivation problems vs. minimizing attrition).

- (5) Design and develop a fully functional and user-friendly research data base that includes all relevant personnel data on 81/82, 83/84, and 86/87 accessions, including all Project A and Career Force Project data and all relevant EMF, Accession File, and Army Training Requirements and Resources System (ATRRS) available data.

Project Organization

To reflect the requirements of research, the project was organized as shown in Figure 1.5. Management of the total project is the responsibility of the Project Director. The overall design, execution, and evaluation of the substantive tasks is the responsibility of the Principal Scientist. Oversight and scientific participation are provided by the Army Research Institute. Guidance is provided by the General Officers Steering Committee and the Scientific Advisory Group. A brief summary of the work encompassed by the three substantive technical tasks is described below.

Task 1 is to revise the measures developed in Project A to measure second-tour soldier performance. The second-tour performance measures will be revised and administered to the Project A Longitudinal Validation sample, beginning in May 1991. At that time, the sample will be in their second tour, and will have been in the Army anywhere from 41 to 63 months. Once these measures have been administered, and the data analyzed (under Task 3), it will be possible to complete the incremental predictive validation of the ASVAB and the Project A Experimental Battery, the measures of training success, and the full array of first-tour performance measures developed in Project A, against the second-tour criterion measures.

Task 2 has a single purpose--to establish, manage, and safeguard an integrated research data base (IRDB) on the National Institutes of Health IBM computer system. As part of the establishment of the IRDB, Task 2 will integrate the Project A longitudinal research data base, extract and merge data from other military data bases, process data collected by Project A and this project, and create workfiles for analyses.

Task 3 is responsible for all analyses performed under this project. The task is organized around the five major data sets to be analyzed. The data sets are the Longitudinal Validation predictor data (LV), the Longitudinal Validation end-of-training data (LV), the Longitudinal Validation first-tour data (LVI), the Concurrent Validation second-tour data (CVII), and the Longitudinal Validation second-tour data (LVII). At the end of the project, Task 3 will have developed the analytic framework necessary to evaluate optimal prediction equations to predict training performance, first-tour performance and attrition, reenlistment, and second-tour performance.

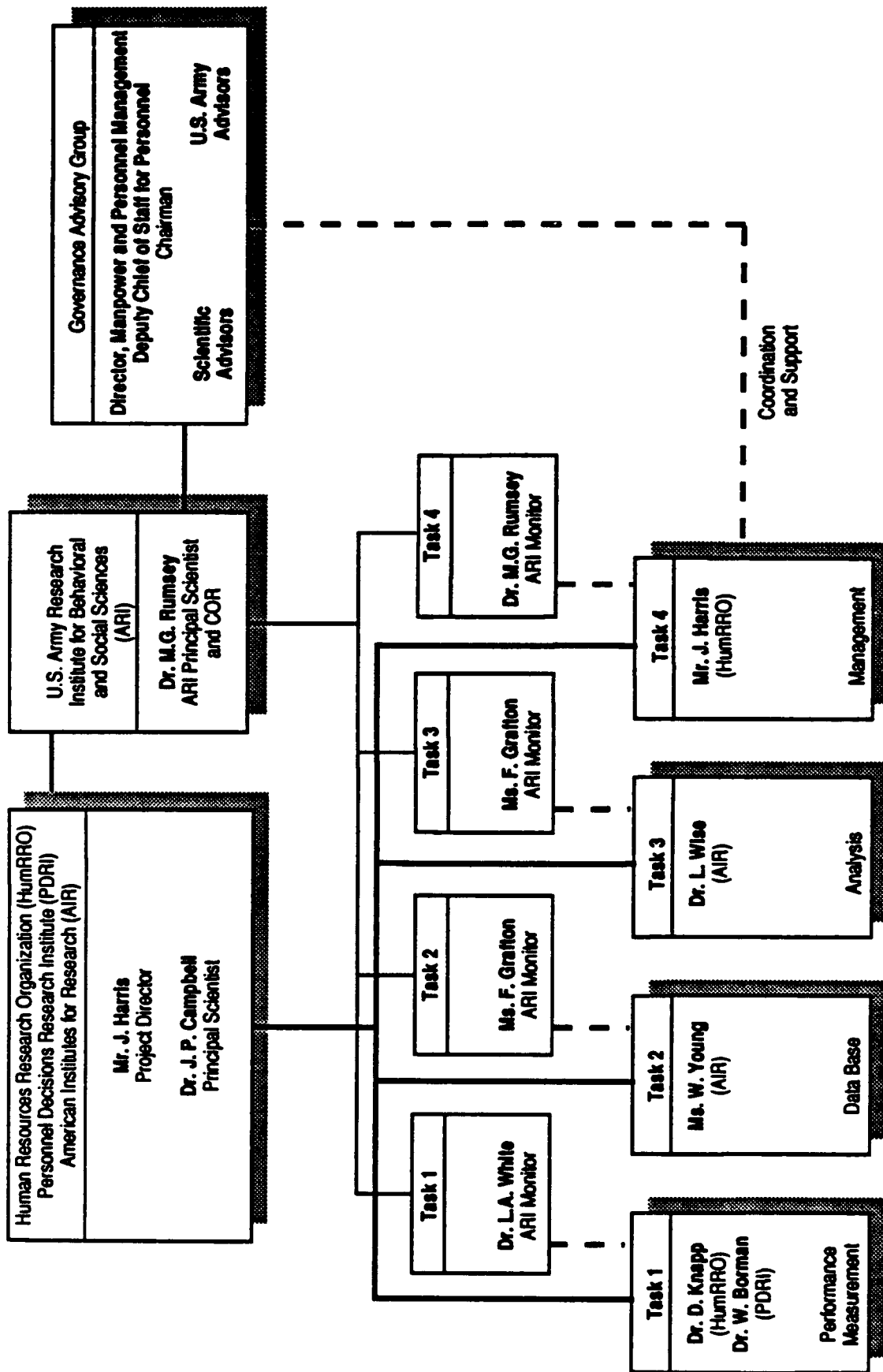


Figure 1.5. Building the Career Force: Initial project management structure.

Summary

The overall design of the Career Force Project includes the development of the remaining parts necessary for a complete modeling of enlisted selection and classification from first tour through second tour. The major design component centers on the collection of job performance criterion data from second-tour job incumbents in 1991 when they are in their 41st to 63rd month of service since enlistment. These individuals entered the Army between 20 August 1986 and 30 November 1987 and were part of the first-tour Longitudinal Validation sample tested as part of Project A. This first-tour sample is referred to as the LVI sample.

We forecast that approximately 100-150 second-tour incumbents per MOS from the LVI sample will be available in 1991. This sample of second-tour soldiers is called the LVII sample.

A major part of the data analytic work for modeling selection and classification will be based on data collected from first-tour soldiers in both the Project A Concurrent Validation (CVI) and Longitudinal Validation (LVI) samples. Since the number of samples for Project A and this project has become rather large, the nomenclature in use is given in Figure 1.6.

The soldiers in the second-tour (LVII) validation sample will be drawn from the same nine MOS originally designated by Project A as the Batch A MOS. They are:

- 11B Infantryman
- 13B Cannon Crewman
- 19K M1 Armor Crewman
- 31C Single Channel Radio Operator
- 63B Light-Wheel Vehicle Mechanic
- 71L Administrative Specialist
- 88M Motor Transport Operator
- 91A Medical Specialist
- 95B Military Police

These are the MOS from which the most criterion data have been collected and for which Project A developed preliminary second-tour NCO job performance measures. However, the original reasons these MOS were chosen for sampling the array of MOS are still relevant. They are high-density MOS that permit reasonable sample sizes and include the highest probabilities for being able to study and account for male/female and black/white differences. They tend to oversample the combat specialties. Within these constraints they represent the total variation in job content across enlisted MOS to the greatest extent possible for a sample of nine MOS.

CONTENT OF THIS REPORT

Chapter 2 is a description of the design and preparation of the integrated research data base, including the Longitudinal Validation predictor sample, the end-of-training data files, the Concurrent Validation second-tour file, and the LV first-tour file. Chapter 3 reports the analysis of the Experimental Predictor Battery, including scoring and forming composites of paper-and-pencil and computer-administered predictors. Chapter 4 presents

Glossary of Terms	
CVI Sample (CVI)	Soldiers who entered the Army between 1 Jul 83 - 30 Jun 84 <u>and</u> were in 1985 Project A Concurrent Validation.
CVII Sample (CVII)	Soldiers who entered the Army between 1 Jul 83 - 30 Jun 84 <u>and</u> were in the 1985 Project A Concurrent Validation (CVI) <u>and</u> the 1988 Second-tour Concurrent Validation (CVII).
LV Sample (LV)	Soldiers in the Longitudinal Validation sample who entered the Army between 20 Aug 86 - 30 Nov 87 <u>and</u> were administered the Experimental Predictor Battery and End-of-Training measures.
LVI Sample (LVI)	Soldiers who entered the Army between 20 Aug 86 - 30 Nov 87 <u>and</u> were in the LV Sample <u>and</u> the 1988 First- Tour Longitudinal Validation Sample.
LVII Sample (LVII)	Soldiers who entered the Army between 20 Aug 86 - 30 Nov 87 <u>and</u> were in the LV Sample <u>and</u> the LVI Sample <u>and</u> who will be in the 1991 Longitudinal Validation (LVII).

Research Flow and Samples

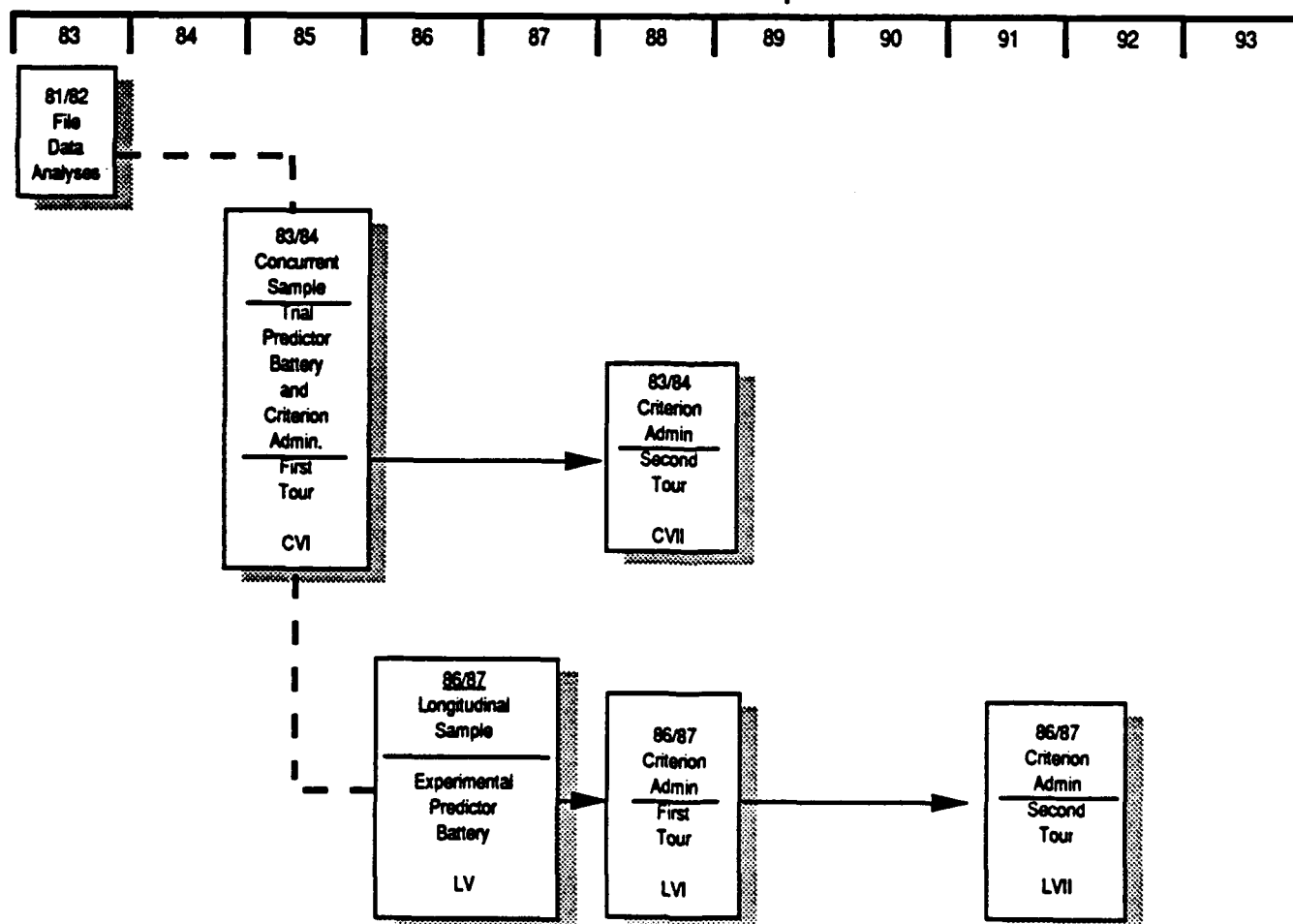


Figure 1.6. Glossary of terms for Project A/Career Force research samples.

results of analyses of the end-of-training measures, including the school knowledge tests and the end-of-training ratings.

Chapter 5 describes the development of scores for second-tour performance measures, including formulating the measures, collecting the second-tour data, and conducting the analysis to determine the basic criterion scores. Chapter 6 is a description of the second-tour performance model, and Chapter 7 presents the plans for the project in the coming years.

Chapter 2 Data File Design and Preparation

Work on the Integrated Research Data Base (IRDB) began with a review of the priority of data needs for the new project as well as the needs of the sponsor. Following consideration of research requirements and establishment of priorities, processing of data began with:

- (1) Longitudinal Validation Predictor sample (LV)
- (2) Longitudinal Validation End-of-Training School Knowledge and Rating data sample (LV)
- (3) Concurrent Validation Second-Tour sample (CVII)
- (4) Longitudinal Validation First-Tour sample (LVI)

The following sections describe each of the data collection efforts, the characteristics of the data collected, sample sizes, and special problems associated with the basic data files.

LONGITUDINAL VALIDATION (LV) PREDICTOR SAMPLE FILES

Project A administered part or all of the Experimental Predictor Battery to 50,235 new recruits between August 1986 and November 1987. Nearly all of the recruits completed the paper-and-pencil battery (six spatial tests, ABLE, AVOICE, and JOB). The computer battery (perceptual and psychomotor) was also administered to nearly all of these recruits, with the exception of the basic infantry MOS (11B). For this MOS, more recruits were available than were required and the computer battery was administered to only one-third of the total.

The distribution of the initial LV predictor sample by MOS is shown in Table 2.1. The sample size ($N = 50,235$) is defined as the number of respondents having either computer data or paper-and-pencil data, or both. Table 2.2 and Table 2.3 show the distribution of the predictor sample by gender and race.

Because resources for processing the predictor data during Project A were limited, processing was discontinued for about 1,000 respondents whose data included SSN errors that made it difficult or impossible to match data from different administrations. Correction of these errors could have resulted in a larger sample for subsequent analyses; however, the cost of making the corrections was not considered to be justified in view of the very large and stable sample that remained. The final sample size for the LV predictor data thus was approximately 49,300.

Table 2.1

Longitudinal Validation (LV) Predictor Sample by MOS: Total Sample

MOS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Unknown	902	1.8	902	1.8
11B Infantryman	14193	28.3	15095	30.0
12B Combat Engineer	2118	4.2	17213	34.3
13B Cannon Crewman	5087	10.1	22300	44.4
16S MANPADS Crewman	800	1.6	23100	46.0
19E M60 Armor Crewman	583	1.2	23683	47.1
19K M1 Armor Crewman	1849	3.7	25532	50.8
27E TOW/Dragon Repairer	139	0.3	25671	51.1
29E Electronics Repairer	257	0.5	25928	51.6
31C Single Channel Radio Operator	1072	2.1	27000	53.7
51B Carpentry/Masonry Specialist	455	0.9	27455	54.7
54E Chemical Operations Specialist	967	1.9	28422	56.6
55B Ammunition Specialist	482	1.0	28904	57.5
63B Light-Wheel Vehicle Mechanic	2241	4.5	31145	62.0
67N Utility Helicopter Repairer	334	0.7	31479	62.7
71L Administrative Specialist	2140	4.3	33619	67.0
76Y Unit Supply Specialist	2756	5.5	36375	72.5
88M Motor Transport Operator	1593	3.2	37968	75.7
91A Medical Specialist	4219	8.4	42187	84.0
94B Food Service Specialist	3522	7.0	45709	91.0
95B Military Police	4206	8.4	49915	99.4
96B Intelligence Analyst	320	0.6	50235	100.0

Table 2.2

LV Predictor Sample by Gender: Total Sample

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Unknown	166	.3	166	.3
Female	5460	10.9	5626	11.2
Male	44609	88.8	50235	100.0

Table 2.3**LV Predictor Sample by Race: Total Sample**

Race	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Unknown	621	1.2	626	1.2
Black	11211	22.3	11832	23.6
Hispanic	1801	3.6	13633	27.1
White	34911	69.5	48544	96.6
Other	1691	3.4	50235	100.0

As mentioned above, the computer battery was not administered to all MOS 11B soldiers in the sample. Table 2.4 shows the sample size separately by MOS for the soldiers who took the computer battery, the paper-and-pencil battery,

Table 2.4**LV Predictor Sample by MOS:
Sample With Both Computer and Paper-and-Pencil Batteries**

MOS	Computer Only	Paper-and- Pencil Only	Both	Total
Unknown	829	16	57	902
11B	0	9653	4540	14193
12B	0	17	2101	2118
13B	0	178	4909	5087
16S	0	17	783	800
19E	0	3	580	583
19K	0	27	1822	1849
27E	0	1	138	139
29E	0	41	216	257
31C	0	102	970	1072
51B	0	13	442	455
54E	0	79	888	967
55B	0	18	464	482
63B	0	120	2121	2241
67N	0	5	329	334
71L	0	196	1944	2140
76Y	0	244	2512	2756
88M	0	53	1540	1593
91A	0	247	3972	4219
94B	0	198	3324	3522
95B	0	81	4125	4206
96B	0	16	304	320
Total	829 (1.7%)	11325 (22.5%)	38081 (75.8%)	50235

or both. The bulk of the difference between the Total Sample (50,235) and the sample having both computer and paper-and-pencil battery (38,081) is explained by the 9,653 11B soldiers who did not take the computer battery. The remaining 2,501 soldiers are scattered throughout the remaining MOS. Reasons for missing a portion of the predictor battery include soldiers being on sick call or other duty. Tables 2.5 and 2.6 show the sample size separately by gender and by race for the soldiers who had both computer and paper-and-pencil data.

Table 2.5

**LV Predictor Sample by Gender:
Sample With Both Computer and Paper-and-Pencil Batteries**

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	4872	12.8	4872	12.8
Male	33209	87.2	38081	100.0

Table 2.6

**LV Predictor Sample by Race:
Sample With Both Computer and Paper-and-Pencil Batteries**

Race	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Unknown	74	.2	74	.2
Black	9353	24.6	9427	24.8
Hispanic	1283	3.4	10710	28.0
White	26086	68.5	36796	96.6
Other	1285	3.4	38081	100.0

Processing of the LV predictor data was done separately for the paper-and-pencil battery and the computer battery, as described below.

Paper-and-Pencil Data Files

For the paper-and-pencil battery, all the data were captured from scanned answer sheets. The sheets were scanned in six batches and the data were stored on magnetic tapes. Each batch of data was processed by first reading in the data separately for each of the measures:

Background Information

6 Cognitive Tests: Orientation Test
 Map Test
 Assembling Objects Test
 Reasoning Test
 Object Rotation Test
 Maze Test

3 Non-Cognitive Tests: ABLE (Assessment of Background
 and Life Experience)
 AVOICE (Army Vocational Interest
 Career Examination)
 JOB (Job Orientation Blank)

The individual measures for each soldier were then combined into one record per soldier. This process revealed several problems, including non-matching due to SSN errors; duplicate matching due to repeated testing, SSN errors, or form code errors; and missing/invalid SSNs. After matching and identification were completed, the processing of the paper-and-pencil data was divided into four steps: (a) read and edit item-level data, (b) score the six cognitive tests, (c) score the three non-cognitive tests, and (d) create composite-level data.

(1) Read and edit item-level data.

Each of the six batches of data was first read and edited separately to resolve any problems unique to each batch. After the data were cleaned within a batch, they were merged with the overall data base. Editing was then done at the more global level. One of the problems encountered most frequently was that of soldiers being tested more than once in different cycles. These cases were resolved by examining the administrator logs and the circumstances surrounding each case.

For this version of the paper-and-pencil file (known as Version I¹), there were 49,408 records and 644 variables (e.g., LV Site Code, SSN Correction Flag, Reasoning Test Items 1-30).

(2) Score the six cognitive tests.

To score the six cognitive tests, key files had to be prepared. A key file contains the correct response to items in a test, and is used to score an individual's performance on that test. The key files were created and checked carefully by both the data base team and the predictor team. Once the key files were verified and the tests were scored, item analysis programs were run to produce item statistics for further analysis. For Version II of the paper-and-pencil file, we added 18 more score variables.

¹The data sets are designated by Version number to indicate how often modifications have been made to the file. "Version I" data sets are the initial reading of the raw data. "Version II" data sets are created from modifications (corrections, creation of new variables, etc.) to the Version I data, etc. All versions are retained, but analyses are performed on only the most recent version.

(3) Score the three non-cognitive tests.

The scoring of the three non-cognitive tests required extensive programming and careful checking of all the variables that went into each scale and also the method by which missing data rules were applied for each test. A total of 47 scale variables were added to Version III of the file.

(4) Create composite-level variables.

The scoring of the composite variables was based on the scoring rules from the Project A CV administration. Different scoring rules may be developed during the course of the present project after project staff have performed additional analyses. At this stage, the scale variables for each of the measures were standardized with a mean of 50 and a standard deviation of 10. The composite variables were created by summing the unit weight of the individual standardized scale variables. This added 17 composite variables to the file, making the final variable count 726.

Computer Battery Data Files

During the data collection, computer battery data were captured on floppy diskettes and sent from the test sites to the data base staff on a weekly basis. The diskettes were logged, processed, and uploaded to an inhouse HP computer each week. The data were then written to magnetic tapes and sent to the mainframe computer for further processing. There were a total of 10 subtests in the computer battery, and the recording system involved 75 records per soldier. The subtests were:

- Cannon Shoot Test
- Target Shoot Test
- Target Tracking 1
- Target Tracking 2
- Short-Term Memory Test
- Perceptual Speed and Accuracy Test
- Target Identification Test
- Number Memory Test
- Simple Reaction Time Test
- Choice Reaction Time Test

The processing of the computer battery data was divided into three steps: (a) read and edit item-level data, (b) create scale-level variables, and (c) create composite-level variables.

(1) Read and edit item-level data.

The item-level data required entering 75 records per soldier. Complications arose when soldiers did not have all 75 expected records. There were two predominant reasons for missing records. First, some soldiers were unable to complete the testing at one sitting and were retested at a later date. Second, there were some problems with the floppy diskettes used to collect the data. Each soldier's data were stored twice, once on a low-density diskette which was sent to the data base manager each week, and once on a high-density diskette which was used as a backup. The problems with data collection

usually occurred on the low-density disks and required using the backup disks to capture the data.

During the processing of item-level data, SSNs for soldiers with incomplete records were sent to project staff for data recovery. In turn, diskettes were sent back containing data recovered from the backup disks. These data were included in the program and checked for further missing data.

This Version I computer battery file had 38,914 records and 1,588 variables.

(2) Create scale-level variables.

The data were cleaned for out-of-range values and "bad" data for the second version of the file. In the next stage, scale-level variables were created from the item data and missing data screens were applied. There were 38,914 observations and 1,915 variables in the Version III file.

(3) Create composite-level variables.

The creation of composite-level variables was based on the scale-level variables that were standardized with a mean of 50 and standard deviation of 10. These scores were then summed by unit weight by battery to create the composites. The method by which these composites were created was the same method used for the CV computer battery.

Due to the size of the file (both number of observations and number of variables), an effort was made to reduce the number of variables by keeping only the scale- and composite-level variables in the final version of the data set.

LONGITUDINAL VALIDATION END-OF-TRAINING DATA FILES

End-of-training (EOT) data were collected on 44,639 soldiers who completed training in the LV predictor MOS between December 1986 and April 1988. Of these, 34,315 (77%) have been matched to the LV predictor sample (soldiers who took the paper-and-pencil and/or computer batteries). Because EOT data were collected by training class, it was not possible to know what proportion of the soldiers in each class had completed the predictor testing until data for the whole class were processed. In addition, some soldiers who took the predictor tests did not complete training during the period when EOT data were collected.

The MOS, gender, and race distribution of the 34,315 EOT records that match the original LV predictor sample are shown in Tables 2.7, 2.8, and 2.9.

EOT data consist of two parts: school knowledge tests and Army-wide ratings.

Table 2.7

**LV End-of-Training (EOT) Sample by MOS:
Cases Matched With LV Predictor Sample**

<u>MOS</u>	<u>Predictor Frequency</u>	<u>EOT Frequency^a</u>	<u>Percent</u>
11B	14193	8117	57.2
12B	2118	1872	88.4
13B	5087	4712	92.6
16S	800	585	73.1
19E	583	442	75.8
19K	1849	1606	86.9
27E	139	92	66.2
29E	257	138	53.7
31C	1072	667	62.2
51B	455	353	77.6
54E	967	616	63.7
55B	482	389	80.7
63B	2241	1215	54.2
67N	334	233	69.8
71L	2140	1414	66.1
76Y	2756	1651	59.9
88M	1593	1354	85.0
91A	4219	3218	76.3
94B	3522	1806	51.3
95B	4206	3639	86.5
96B	320	196	61.3
Total	49333 ^b	34315	69.6

^a Soldiers who took the paper-and-pencil and/or computer batteries, and who are also in the LV End-of-Training sample.

^b Does not include 902 soldiers in unknown MOS (see Table 2.1).

Table 2.8

**LV End-of-Training (EOT) Sample By Gender:
Cases Matched With LV Predictor Sample**

<u>Gender</u>	<u>Predictor Frequency</u>	<u>EOT Frequency^a</u>	<u>Percent</u>
Female	5460	3491	6.9
Male	44609	30824	69.1
Total	50069 ^b	34315	68.5

^a Soldiers who took the paper-and-pencil and/or computer batteries, and who are also in the LV End-of-Training sample.

^b Does not include 166 soldiers whose gender is unknown.

Table 2.9**LV End-of-Training Sample by Race:
Cases Matched With LV Predictor Sample**

<u>Race</u>	<u>Predictor Frequency</u>	<u>EOT Frequency^a</u>	<u>Percent</u>
Unknown	621	31	5.0
Black	11211	7986	71.2
Hispanic	1801	1316	73.1
White	34911	23755	68.0
Other	1691	1227	72.6
Total	50235	34315	68.3

^a Soldiers who took the paper-and-pencil and/or computer batteries, and who are also in the LV End-of-Training sample.

School Knowledge Test Data Files

A total of 44,392 of the 44,639 EOT soldiers took the school knowledge tests on machine scannable answer sheets. These answer sheets were scanned and sent to the data base manager in the form of magnetic tapes in six separate batches. The data were processed in three steps: (a) read and edit item-level data; (b) score tests and create scale-level data; (c) establish missing data and random response flags, and create composite-level data.

(1) Read and edit item-level data.

Because each MOS had a different school knowledge test, the data were processed separately for each of the 21 MOS by first linking the school knowledge test records with the Background Information sheets to determine the MOS code. The MOS code was not physically printed on the school knowledge test answer sheets because a generic answer sheet was used. This caused a problem for the processing of the MOS 13B (Cannon Crewmen) answer sheets. MOS 13B has two tracks (self-propelled and towed) and soldiers in each track took different school knowledge tests based on their track. However, on the Background Information sheets, the MOS code was identified only as 13B. After conferences with the overall test site coordinator, corrective actions were taken ensuring that all but seven 13B cases were captured (out of a total of 5,281).

Twenty-two Version I data files, one per MOS (two for 13B), were created, as follows:

<u>MOS</u>	<u>Number of Cases</u>	<u>Number of Variables</u>
11B	10,575	148
12B	2,001	160
13S	4,356	165 (13B self-propelled)
13T	925	166 (13B towed)
16S	694	153
19E	471	167
19K	1,659	165
27E	166	173
29E	306	155
31C	1,377	167
51B	377	171
54E	808	160
55B	674	185
63B	1,451	163
67N	408	165
71L	1,843	102
76Y	2,289	170
88M	1,913	135
91A	5,368	172
94B	2,695	132
95B	3,776	125
96B	253	166

(2) Score tests and create scale-level data.

Before scoring each test, we prepared 22 key files to score the individual school knowledge tests (one per MOS, and two for 13B). One problem with the key files resulted from the late return of the proponent reviews, which required changing the test scoring routines after the tests were already in the field. Several MOS, such as 63B and 71L, were affected and special changes in the programs were necessary in order for the tests to be scored correctly.

Separate item analyses were run and checked for each of the MOS to ensure accuracy in scoring programs. Five variables were created for each scale (i.e., task test) and stored in the Version II data sets for each MOS.

(3) Establish missing data and random response flags, and create composite-level data.

The Version III school knowledge test data sets included three procedures:

- Set missing data flag
- Set random response flag
- Create composite-level variables

The missing data flag was set when more than 10 percent of an individual's overall test items were missing. The random response flag was set on the basis of a random response index defined as the correlation between the item score (1 for correct and 0 for incorrect) and the item difficulty (expressed as the proportion of subjects who answered the item correctly). For most individuals the value for this correlation was positive, but in a few cases it was essentially zero, suggesting the soldier was responding randomly. For individuals flagged either for missing data or for random response, the composite-level data were shown as missing.

Three composite-level variables were created for each MOS. The basis for forming the composite scores is given in Chapter 4.

End-of-Training Ratings Data Files

The processing of the EOT Army-wide ratings was more time consuming than anticipated. During the EOT testing, peer and supervisor ratings were collected for each soldier. On the ratings answer sheets, soldiers' three-digit ID numbers were entered on the sheets but their class IDs were not. Since the three-digit soldier ID was not unique², class IDs were obtained from the Background Information Sheets that were filled out by the soldiers or the supervisors. Several problems were encountered during the processing of the rating forms: missing class ID and/or soldier ID, duplicate class ID and/or soldier ID, incomplete class ID and/or soldier ID, or invalid class ID and/or soldier ID.

Due to limited resources, many of these rating problems were not resolved during Project A. We estimated that these problems might account for as many as 30 percent of the total ratings (i.e., rater-ratee pairs) that were collected. Tables 2.10, 2.11, and 2.12 provide MOS, gender, and race information for the rater-ratee pair information that we were able to process. Since many trainees had multiple raters, the numbers in the tables do not correspond to the number of people in the sample. When the school knowledge test data were matched with the EOT ratings, all but 639 of the 44,392 soldiers with a knowledge test score had at least one rating.

The processing of the rating data was done in three steps: (a) read and edit rater-ratee pair data; (b) create one record per ratee file, after performing outlier and missing data analyses; (c) create composite-level data.

(1) Read and edit rater-ratee pair data.

The rating records were read as rater-ratee pairs. Each record was matched against the link file (the master file of soldiers to be included in the analyses) by post code, class number, and the soldier ID for verification. Any records that we were not able to verify against the link file (and therefore were not able to obtain an SSN for matching up other testing records) were flagged as deletes and removed. An initial effort was made to correct such problems, but due to limited resources during Project A the efforts were not continued.

²Soldier IDs were assigned by MOS and site. Each ID was unique within site, but the same ID could be used at multiple sites.

Table 2.10

LV End-of-Training Rater-Ratee Sample by MOS^a

MOS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Unknown	1	.0	1	.0
11B	8267	18.8	8268	18.8
11C	807	1.8	9075	20.6
11H	5	0.0	9080	20.6
11M	529	1.2	9609	21.8
11X	684	1.6	10293	23.4
12B	1996	4.5	12289	27.9
13B	5287	12.0	17576	40.0
16S	691	1.6	18267	41.5
19E	481	1.1	18748	42.6
19K	1655	3.8	20403	46.4
27E	179	0.4	20582	46.8
29E	306	0.7	20888	47.5
31C	1358	3.1	22246	50.6
51B	387	0.9	22633	51.5
54E	805	1.8	23438	53.3
55B	689	1.6	24127	54.9
63B	1458	3.3	25585	58.2
67N	407	0.9	25992	59.1
71L	1831	4.2	27823	63.3
76Y	2253	5.1	30076	68.4
88M	1921	4.4	31997	72.8
91A	5334	12.1	37331	84.9
94B	2631	6.0	39962	90.9
95B	3766	8.6	43728	99.4
96B	250	0.6	43978	100.0

^a Since many trainees had multiple raters, the numbers in the table do not correspond to the number of people in the sample.

Table 2.11

LV End-of-Training Rater-Ratee Sample by Gender^a

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Unknown	419	1.0	419	1.0
Female	4636	10.5	5055	11.5
Male	38923	88.5	43978	100.0

^a Since many trainees had multiple raters, the numbers in the table do not correspond to the number of people in the sample.

Table 2.12**LV End-of-Training Rater-Ratee Sample by Race^a**

<u>Race</u>	<u>Frequency</u>	<u>Percent</u>	<u>Cumulative Frequency</u>	<u>Cumulative Percent</u>
Unknown	453	1.0	453	1.0
Black	10276	23.4	10729	24.4
Hispanic	1714	3.9	12443	28.3
White	29891	68.0	42334	96.3
Other	1644	3.7	43978	100.0

^a Since many trainees had multiple raters, the numbers in the table do not correspond to the number of people in the sample.

The Version I file was created with 193,932 observations (rater-ratee pairs) and 20 variables.

- (2) Perform outlier and missing data analyses, and create one record per ratee file.

Before creating one record per ratee, rater-ratee pair records were deleted for being outliers or for missing too many ratings. Records were then combined to form one record per ratee. Within each record, an average peer rating, an average supervisor rating, and a combined peer/supervisor rating were computed. The Version II file had 43,978 observations (ratees) and 24 variables.

- (3) Create composite-level data.

Composite-level variables were created by taking the average of the ratings that were used in that composite. The ratings were not standardized because the standard deviations did not differ sufficiently to cause concern about weighting the separate scales. Twelve composite rating variables (described in Chapter 4) were added to the Version III file.

CONCURRENT VALIDATION SECOND-TOUR (CVII) FILES

Second-tour performance data were collected from 1,053 soldiers between July 1988 and February 1989. Only Batch A MOS personnel were assessed. Despite the effort that was devoted in the field to capturing second-tour soldiers who were also in our first-tour CV sample, only 163 of the 1,053 CVII soldiers were in the original first-tour sample.

The MOS, gender, and race composition of the CV second-tour file are shown in Tables 2.13, 2.14, and 2.15.

Table 2.13

Concurrent Validation (CVII) Sample by MOS

<u>MOS</u>	<u>Frequency</u>	<u>Percent</u>	<u>Cumulative Frequency</u>	<u>Cumulative Percent</u>
11B	127	12.1	127	12.1
13B	162	15.4	289	27.4
19E	33	3.1	322	30.6
19K	10	0.9	332	31.5
31C	103	9.8	435	41.3
63B	116	11.0	551	52.3
71L	112	10.6	663	63.0
88M	144	13.7	807	76.6
91A	105	10.0	912	86.6
95B	141	13.4	1053	100.0

Table 2.14

CVII Sample by Gender

<u>Gender</u>	<u>Frequency</u>	<u>Percent</u>	<u>Cumulative Frequency</u>	<u>Cumulative Percent</u>
Female	114	10.8	114	10.8
Male	939	89.2	1053	100.0

Table 2.15

CVII Sample by Race

<u>Race</u>	<u>Frequency</u>	<u>Percent</u>	<u>Cumulative Frequency</u>	<u>Cumulative Percent</u>
Black	347	33.0	347	33.0
Hispanic	70	6.6	417	39.6
White	580	55.1	997	94.7
Other	56	5.3	1053	100.0

CVII data consisted mainly of criterion data with the exception of ABLE testing. ABLE testing was given only when soldiers were not able to participate in the rating session. As briefly described in Chapter 1, the instruments that were used in the second-tour testing were as follows:

- Hands-On Tests
- Job History Questionnaire
- Job Knowledge Tests
- Army-Wide Ratings
- MOS-Specific Ratings
- Combat Performance Prediction Scale
- Personnel File Form
- Army Job Satisfaction Questionnaire
- Role Play
- Situational Judgment Test
- Measurement Method Ratings
- ABLE

In the following sections, the processing of each of the instruments will be discussed. The basic scores that were developed for each instrument and the procedures used to identify them are described in Chapter 5 of this report.

Hands-On Tests. Hands-on tests were designed to examine the soldier's ability to perform specific tasks by having the soldier perform the task. The data were processed separately for each MOS. Before the data were keytaped, each hands-on scoresheet was checked by a data clerk to ensure the accuracy of information on the scoresheet, such as soldier ID, scorer ID, post ID, and information on each hands-on step. For certain MOS (such as 71L and 91A), rescoring of some scoresheets was necessary to correct original scorer mistakes or to indicate some circumstance affecting testing, such as lack of equipment.

After the scoresheets for each individual task were keytaped, all tasks for individual soldiers were combined into one record per soldier. A matching process was used to ensure that records for each soldier were linked correctly. Two versions of the files were created. Version I edited each step variable and merged all individual records. Version II performed missing data analyses and created task-level variables by combining the step variables to form a total percent GO score for each task. The following represents basic file information for the Version II files:

<u>MOS</u>	<u>Number of Observations</u>	<u>Number of Variables</u>
11B	126	242
13B	142	532
19E	41	232
31C	96	366
63B	112	238
71L	109	293
88M	136	301
91A	93	386
95B	139	377

During CVII hands-on testing, shadow scoring--that is, scoring by a second team member for comparison purposes--data were also collected for two MOS, 11B and 91A. The processing of the shadow scoring data followed the same rules as those used on the primary scoring data described above. Two files were created:

<u>MOS</u>	<u>Number of Observations</u>	<u>Number of Variables</u>
11B	23	242
91A	37	386

Job History Questionnaire. Job History Questionnaire data were also processed separately by MOS. Each MOS file contained variables indicating "how often" and "how recent" a soldier performed certain tasks. Version I files were created to read and edit item-level data, and also created two summary variables. The basic file information is as follows:

<u>MOS</u>	<u>Number of Observations</u>	<u>Number of Variables</u>
11B	129	63
13B	159	67
19E	43	57
31C	103	59
63B	117	59
71L	112	65
88M	144	61
91A	105	59
95B	145	65

Job Knowledge Tests. Job knowledge tests were MOS-specific, multiple-choice tests of job proficiency and were processed separately by MOS. For each MOS, a key file was established for scoring and item analysis purposes. Two versions of the file were created. Version I read and edited item-level variables, and Version II scored items and created composite-level variables.

<u>MOS</u>	<u>Number of Observations</u>	<u>Number of Variables</u>
11B	129	432
13B	161	422
19E	42	346
31C	103	325
63B	116	295
71L	112	354
88M	144	342
91A	105	332
95B	146	316

Army-Wide Ratings. Verifying the ratee ID against the link file and locating missing/ invalid ID information took a great deal of time. Three versions were created for these data. The first version read in rater-ratee pair information and merged it with the link file to identify ID errors. Version II corrected ID errors and deleted records that could not be corrected. Finally, Version III consolidated all rating records (rater-ratee pairs) into one record per ratee and created scale- and composite-level variables. This file had 962 observations and 79 variables.

MOS-Specific Ratings. MOS-specific ratings, while different for each MOS, were processed in one file because the differences were in the content of the various ratings rather than in the format of the data. Two versions of the file were created. Version I read in the rater-ratee pair rating information and merged in ratee ID information from the Army-wide ratings file. Discrepancies between ratee ID information from the Army-wide data and ratings information from the MOS-specific file were resolved. These discrepancies usually involved individuals who received Army-wide ratings but did not receive MOS-specific ratings. Version II produced one record per ratee, and created scale- and composite-level variables. This file had 948 observations and 53 variables.

Combat Performance Prediction Scale. The processing of the combat ratings was very similar to that of the MOS-specific ratings. One additional edit dealt with female soldiers. During CVII testing, the policy was to collect combat ratings for male soldiers only. Even though most female soldiers did not have combat ratings, a few did and these were deleted after verifying their gender against the link file. Two versions were created for this file. Version I read and edited rater-ratee pair level data. Version II created one record per ratee, and created scale- and composite-level variables. This file had 854 observations and 67 variables.

Personnel File Form. Personnel File Form data used four versions of the file. Version I read in data and performed basic edits. Version II created several scale-level variables. Version III added awards variables that were not captured originally. Version IV created additional scale variables and composite variables. This file had 1,060 individuals and 90 variables.

Army Job Satisfaction Questionnaire. The Army Job Satisfaction Questionnaire (AJSQ) data were scanned along with all the LV first-tour data. The item-level data were edited together with the first-tour data to create a Version I file. The Version II file created scale- and composite-level data. The file had 1,014 observations and 43 variables.

Situational Judgment Test. The Situational Judgment Test asked soldiers to indicate both the most likely and the least likely response to situations. The editing of this test was time consuming due to the different options available. Editing occurred in two stages. In stage one, data clerks edited the raw data by hand before the tests were sent to keytaping. In stage two, further editing was performed in creating the Version I data set. Two versions of the file were created. Version I read in and edited item data, and Version II created composite-level variables. This file had 1,048 observations and 182 variables.

Simulation (Role Play Exercises). The processing of the role play data was similar to that for the hands-on data. Role play data consisted of three rating score sheets:

- Checklist of Disciplinary Counseling Behavior
- Checklist of Personal Counseling Behavior
- Checklist of Training Behavior

Each score sheet was checked for soldier ID, post ID, and scorer ID before they were merged into one record per soldier. In addition to primary scoring, shadow scoring was also collected during the testing. Unlike the

hands-on data, however, both primary and shadow scoring data were combined in one file instead of separate files. Two versions of the file were created. Version I read in each score sheet, performed basic editing, and then merged all records for each soldier. Version II created composite-level variables. This file contained 979 observations and 111 variables.

Measurement Method Ratings. Measurement Method ratings were collected at the end of the testing to determine perceptions of fairness for each measure. Each MOS had slightly different ratings. Nine files were created, one for each MOS. Only Version I files were created.

<u>MOS</u>	<u>Number of Observations</u>	<u>Number of Variables</u>
11B	111	21
13B	103	21
19E	40	21
31C	38	21
63B	74	21
71L	58	21
88M	111	21
91A	56	21
95B	94	21

ABLE. Data for ABLE were scanned along with the first-tour data. These data were processed using the same rules as were used for the LV predictor data. Two versions of the file were created. Version I read in and edited ABLE items, and Version II computed the scale and composite variables. This file had 691 observations and 257 variables.

LONGITUDINAL VALIDATION FIRST-TOUR (LVI) DATA

First-tour performance data were collected on 11,266 soldiers in 21 MOS between July 1988 and February 1989. As with the second-tour data, only criterion data were collected, with the exception of the retesting of ABLE.

The MOS, gender, and race composition of the LV first-tour file are shown in Tables 2.16, 2.17, and 2.18.

Table 2.16**LVI Sample by MOS**

<u>MOS</u>	<u>Frequency</u>	<u>Percent</u>	<u>Cumulative Frequency</u>	<u>Cumulative Percent</u>
11B	909	8.1	909	8.1
12B	841	7.5	1750	15.5
13B	916	8.1	2666	23.7
16S	472	4.2	3138	27.9
19E	249	2.2	3387	30.1
19K	824	7.3	4211	37.4
27E	90	0.8	4301	38.2
29E	112	1.0	4413	39.2
31C	529	4.7	4942	43.9
51B	213	1.9	5155	45.8
54B	499	4.4	5654	50.2
55B	279	2.5	5933	52.7
63B	752	6.7	6685	59.3
67N	197	1.7	6882	61.1
71L	678	6.0	7560	67.1
76Y	788	7.0	8348	74.1
88M	682	6.1	9030	80.2
91A	824	7.3	9854	87.5
94B	832	7.4	10686	94.9
95B	452	4.0	11138	98.9
96B	128	1.1	11266	100.0

Table 2.17**LVI Sample by Gender**

<u>Gender</u>	<u>Frequency</u>	<u>Percent</u>	<u>Cumulative Frequency</u>	<u>Cumulative Percent</u>
Female	1317	11.7	1317	11.7
Male	9949	88.3	11266	100.0

Table 2.18

LVI Sample by Race

<u>Race</u>	<u>Frequency</u>	<u>Percent</u>	<u>Cumulative Frequency</u>	<u>Cumulative Percent</u>
Black	3090	27.4	3090	27.4
Hispanic	405	3.6	3495	31.0
White	7346	65.2	10841	96.2
Other	425	3.8	11266	100.0

As mentioned above, only criterion measures were administered to the LVI sample, with the exception of the ABLE, which was administered to soldiers who were unable to rate other soldiers. A list of the instruments is as follows:

	<u>Batch A MOS</u>	<u>Batch Z MOS</u>
Hands-On Tests	X	
Job History Questionnaire	X	
Job Knowledge Tests	X	
Army-Wide Ratings	X	X
MOS-Specific Ratings	X	
Combat Performance Prediction Scale	X	X
Personnel File Form	X	X
Army Job Satisfaction Questionnaire	X	X
School Knowledge Tests (Batch Z MOS only)		X
ABLE	X	X

In the following sections, the processing of each of the instruments will be discussed. The development of basic scores for the LVI performance measures is still in progress. The analytic procedures that were used and the results that were obtained will be described in subsequent reports.

Hands-On Tests. The hands-on tests that had been designed for the CV sample were used again for the LVI sample. However, some tasks were eliminated because (a) they were no longer part of the soldiers' duties, (b) they were no longer taught at OSUT/AIT, or (c) the Army requested they be dropped. In addition, one task, Engage Target with M16, was added to the MOS 11B and 95B tasks.

The data were processed separately for each MOS. Before the data were keytaped, a data clerk checked each hands-on scoresheet to ensure the accuracy of information on the scoresheet, such as soldier ID, scorer ID, post ID, and information on each hands-on step. Rescoring was necessary for some tests within each MOS. In some cases the rescoring simply indicated that at a particular site a piece of equipment necessary to the test was not available and so the step was not applicable to that site. In other cases, the entire task had to be rescored because of original scorer mistakes.

After the scoresheets for each task were keytaped, all tasks for individual soldiers were combined into one record per soldier by MOS. A matching process was used to ensure that records for each soldier were linked correctly. The following represents basic file information for the LVI data:

<u>MOS</u>	<u>Number of Tasks</u>
11B	14
13B	16
19E	15
19K	14
31C	15
63B	13
71L	13
88M	15
91A	13
95B	13

As with the CVII hands-on testing, shadow scoring data were collected for two MOS, 11B and 91A. The processing of the shadow scoring data followed the same rules used on the primary scoring data previously described.

Job History Questionnaire. Job History Questionnaire data were processed separately by MOS. Each MOS file contained variables indicating "how often" and "how recent" a soldier performed certain tasks. The basic file information is as follows:

<u>MOS</u>	<u>Number of Observations</u>	<u>Number of Variables</u>
11B	905	65
13B	916	65
19E	251	65
19K	821	63
31C	498	65
63B	750	63
71L	665	57
88M	680	65
91A	817	63
95B	452	67

Job Knowledge Tests. The Job Knowledge tests used for the CV data collection were used again for the LVI collection. These tests were MOS-specific, multiple-choice tests of job proficiency and were processed separately by MOS. For each MOS, a key file was established for scoring and item analysis purposes. Following is basic file information for the LVI data:

<u>MOS</u>	<u>Number of Tasks</u>
11B	28
13B	30
19E	29
19K	30
31C	30
63B	29
71L	24
88M	28
91A	29
95B	30

Army-Wide Ratings. The LVI Army-wide ratings data were collected using scannable answer booklets. Because of the large volume of booklets, after scanning the booklets themselves were stored by the scanning company. Without the original documents for verification, processing and locating missing/invalid ID information was a lengthy operation. After processing, 45,058 of the 48,751 rater/ratee pairs in the raw data (92%) were available for further analysis.

MOS-Specific Ratings. MOS-specific ratings had to be processed simultaneously with the Army-wide ratings because the only source of ratee and rater ID information was on the Army-wide ratings. The MOS-specific ratings, while different for each MOS, were processed in one file because the differences in the ratings are in the content of the ratings rather than in the format of the data. MOS-specific ratings were collected for Batch A MOS only. After processing, 26,473 of the 28,334 rater/ratee pairs in the raw data (93%) were available for further analysis.

Combat Performance Prediction Scale. Processing of the combat ratings was also done simultaneously with the Army-wide ratings because of the ID information. In addition, any combat ratings collected on female soldiers needed to be eliminated because the policy was to collect combat ratings for male soldiers only. After processing, 40,404 of the 45,152 rater/ratee pairs in the raw data (89%) were available for further analysis.

Personnel File Form. Personnel File Form data were collected for 11,237 soldiers. These data represent self-reported administrative data including awards received, memoranda/letters of appreciation received, certificates of appreciation received, Physical Readiness Test Score, M16 Qualification, Skill Qualification Test³ score, Articles 15 received, and Flag Actions received.

Army Job Satisfaction Questionnaire. Data regarding job satisfaction were collected from 12,265 soldiers. The data include job satisfaction dimensions relevant to the Army and information about the soldiers' background.

³A criterion-referenced, paper-and-pencil performance-knowledge test which evaluated the soldier's ability to perform critical job tasks.

School Knowledge Tests. The School Knowledge test data were collected from the Batch Z MOS only. These MOS-specific training achievement tests were processed separately by MOS using the same procedures used for the EOT School Knowledge tests. These were administered to Batch Z soldiers to supplement performance data collected from them. Recall that Batch Z soldiers have no hands-on, job knowledge, or MOS-specific ratings data.

ABLE. ABLE data were collected from 1,272 soldiers who were unable to rate other soldiers.

UPDATES FROM EXISTING ARMY DATA

The Career Force data base contains a great deal of information extracted from operational military data files, in addition to the data collected by project staff. These data include information on operational selection and classification test scores, progress during and after training, the timing and nature of separations, and additional proficiency or performance information.

The operational files included in the Career Force data base are:

- Applicant/Accessions Data
- Training Data
- Skill Qualification Test Data
- Enlisted Master File Data
- Defense Manpower Data Center Cohort Data
- World-Wide Locator Data

Applicant/Accessions Data. We currently have applicant/accessions data from 1981 to 1988, which have been processed on an annual basis. However, because of major changes in the applicant/accessions formats, and the fact that the Project plans do not call for collecting data on any new cohort (i.e., beyond 1988), the processing of applicant/accessions is likely to be put on a lower priority basis.

Training Data. For certain periods, we processed training data received from the Army Training and Doctrine Command (TRADOC). Currently, the Career Force data base includes training data for FY81/82, FY83, and FY88.

Skill Qualification Test Data. Since 1983, we have collected SQT data from the Soldier Support Center at Fort Eustis. These data were generally processed on an annual basis.

Enlisted Master File Data. EMF data have been processed on a quarterly basis beginning in 1984. These data provide current information on the status of each soldier in our analysis cohorts including, in particular, current paygrade, reenlistment status, and separation status. Since the collection of the LV predictor sample, reenlistment and separation information has been regularly compiled and added to each sample link file for analysis.

Defense Manpower Data Center Cohort Data. While the EMF is a "current status" system, the DMDC cohort data provide a historical file for each accessions cohort. These files contain more complete historical information based on EMF transaction data. Since we have relied more heavily on the EMF

for reenlistment/separation data, the processing of the DMDC data has become a lower priority.

World-Wide Locator Data. We process World-Wide Locator data from Fort Benjamin Harrison on an "as needed" basis to obtain the most up-to-date location information on our sample cohort. Most recently, we have been receiving quarterly data from Fort Benjamin Harrison for our Longitudinal samples to be prepared for our second-tour testing in the summer of 1991.

GENERATION OF CAREER FORCE PROJECT WORKFILES

A workfile is defined as any file that is used by project staff members (or other researchers) who are not directly associated with the data base. Each researcher generally requires different variables and workfiles contain only those variables needed by the individual researcher. Access to these files is given on an "as-needed" basis.

To maintain control of the use of data from the Career Force Integrated Research Data Base (IRDB), an Online Request Form has been developed. Each potential workfile user is required to provide information indicating what kind of data is needed, samples for which the data are needed, why the data are needed, and how the data will be used. The IRDB manager reviews the request to make sure that the data being requested are available and that no similar workfiles have already been requested by another user. The request is then forwarded to the Project Director and the ARI COR for formal approval.

After the request has been approved, the programs needed to create the required workfile are prepared. Each workfile has an individual file name that is connected to a data collection phase and followed by a sequential version identifier. Each workfile contains an encrypted ID and the relevant variables, and is RACF protected so that only the researcher, the COR, and the IRDB manager have access to the file. With the approval of the Project Director and the COR, other users may have access to the workfile if they share similar research interests.

Workfiles are monitored quarterly, and files that have not been used within the quarter are backed up to tape. Workfiles can also be updated when new information is available; however, this requires an additional workfile request that must be approved by the Project Director and the COR.

Chapter 3

Analysis of the Experimental Predictor Battery: LV Sample

INTRODUCTION

The development of the Experimental Predictor Battery has been an iterative process. It has utilized data from three cycles: the Pilot Trial Battery (PTB) administered in the field test of Project A (Peterson, 1987), the Trial Battery (TB) administered in the Concurrent Validation of Project A (Campbell, 1990), and the Experimental Predictor Battery (EB) administered as part of the Longitudinal Validation effort, and the subject of this chapter.

One principle has been to build upon the knowledge gained in past iterations to improve both the measures in the battery and the methods for using the battery, but also to refrain from making unnecessary modifications. The predictor measures were used successfully in the large-scale Concurrent Validation (McHenry, et al., 1990; Peterson, et al., 1990). Any changes must be made in light of this experience.

A detailed analysis of the Experimental Battery is necessary for several specific reasons. Some of the measures were revised after the Concurrent Validation analyses, turning the Trial Battery into the Experimental Battery. The effect of these changes must be evaluated. Also, the Experimental Battery was administered to a much larger and different type of sample than was the Trial Battery. Whereas the Trial Battery was administered to about 10,000 soldiers who had completed basic and advanced training and had been on the job for about 12-24 months, the Experimental Battery was administered to about 50,000 soldiers who had just entered the Army (they were tested during the first three days of their enlistment). Both samples are large enough to produce stable results, but there are major differences in terms of their experiences with Army life and its possible effects on performance on the measures in the predictor battery. Finally, the project scientists and advisors were able to make use of additional results obtained in earlier research with the Trial Battery, both from Project A and from other research projects conducted by the Army Research Institute.

The major technical goals for the analysis of the Experimental Battery remained the same as for earlier analyses of the Trial Battery: (a) utilize appropriate screening steps to ensure the quality of the data, (b) determine the most appropriate method to compute basic scores on the tests, (c) report the basic descriptive statistics and psychometric properties of the basic test scores, and (d) perform appropriate analyses to recommend composite scores.

This last step is important because there are more than 60 basic test scores in the Experimental Battery. Entering such a large number of scores into multiple regression equations to predict job performance criteria presents problems, especially for Army jobs with relatively small sample sizes. It is, therefore, highly desirable to reduce the number of scores to be used to predict job performance. On the other hand, each measure had been included because it was deemed important for predicting job performance and it is equally important to preserve, as much as possible, the specificity or heterogeneity of the original set of test scores. The challenge is to balance these competing priorities.

The Experimental Battery

The Experimental Battery (EB) includes three major types of instruments: (a) cognitive paper-and-pencil tests designed to measure spatial constructs; (b) computer-administered tests of cognitive, perceptual, and psychomotor abilities; and (c) non-cognitive paper-and-pencil measures of temperament, interests, and job preference. Names of the instruments and the constructs they were designed to measure appear in Figure 3.1. Peterson (1987) and Peterson, et al. (1990) provide complete descriptions of the constructs and test development. Appendix A provides descriptions and sample items for the cognitive paper-and-pencil and computer-administered tests, and Appendix B provides definitions for the elements in the three non-cognitive inventories. Working paper documentation describing the computer tests and analyses, the LV scoring program, and the paper-and-pencil tests and analyses in detail were supplied to ARI in October 1990.¹

This chapter is divided into eight major sections. This introductory section describes the overall approach and sampling strategy. The second section describes analyses of the paper-and-pencil spatial predictors. The third section summarizes analyses of the computer-administered tests, and the fourth describes analyses including all cognitive predictor composites. The next three sections describe analyses of the three non-cognitive predictors: ABLE, AVOICE, and JOB respectively. The final section is a brief summary listing the recommended predictor composites for use in the Career Force Project validation analyses.

The Analysis Samples

Of the 50,235 soldiers who were tested during the longitudinal data collection, 38,081 had complete predictor data (i.e., computer-administered and paper-and-pencil predictor data). We obtained paper-and-pencil predictor data only (i.e., no computer-administered predictor data) for 11,325 soldiers, mostly in combat MOS, particularly 11B Infantrymen.

To conserve computing resources, most analyses were conducted on a sample of 7,000 soldiers with complete predictor data. This sample, called the "Initial Sample," is a random sample stratified on race, gender, and MOS. Confirmatory analyses were conducted on a second random sample stratified on race, gender, and MOS (N = 7,000), called "Sample 2." Both subsamples were drawn from the sample of incumbents having complete predictor data (i.e., about 38,000). Although the basic N for each subsample was 7,000, the numbers vary slightly for the various analyses, as noted in the tables that follow in this chapter.

Table 3.1 provides the numbers of soldiers in each MOS in the total sample compared to the subsamples. Table 3.2 provides breakdowns of the sub- and total samples by race and gender. As shown, the demographics of the total sample are closely reflected by both the subsamples. The Ns shown in these tables are slightly lower than some shown in Chapter 2 summary tables because

¹As this documentation includes sensitive information on scoring the various measures, it is accessible only through special permission from the Chief, Selection and Classification Technical Area.

Test/Measure	Construct
<i>Paper-and-Pencil Spatial Tests</i>	
Assembling Objects	Spatial Visualization-Rotation
Object Rotation	Spatial Visualization-Rotation
Maze	Spatial Visualization-Scanning
Orientation	Spatial Orientation
Map	Spatial Orientation
Reasoning	Induction
<i>Computer-Administered Tests</i>	
Simple Reaction Time	Reaction Time (Processing Efficiency)
Choice Reaction Time	Reaction Time (Processing Efficiency)
Short-Term Memory	Short-Term Memory
Perceptual Speed and Accuracy	Perceptual Speed and Accuracy
Target Identification	Perceptual Speed and Accuracy
Target Tracking 1	Psychomotor Precision
Target Shoot	Psychomotor Precision
Target Tracking 2	Multilimb Coordination
Number Memory	Number Operations
Cannon Shoot	Movement Judgment
<i>Temperament, Interest, and Job Preference Measures</i>	
Assessment of Background and Life Experiences (ABLE)	Adjustment Dependability Achievement Physical Condition Leadership (Potency) Locus of Control Agreeableness/Likability
Army Vocational Interest Career Examination (AVOICE)	Realistic Interest Conventional Interest Social Interest Investigative Interest Enterprising Interest Artistic Interest
Job Orientation Blank (JOB)	Job Security Serving Others Autonomy Routine Work Ambition/Achievement

Figure 3.1. Experimental Predictor Battery tests and relevant constructs.

Table 3.1

Longitudinal Validation: Comparison of Initial Sample and Sample 2 Demographics to the Complete Sample by MOS

MOS	Complete Sample ^a		Initial Sample		Sample 2	
	N	Percent	N	Percent	N	Percent
11B Infantryman	4540	11.92	834	11.93	835	11.95
12B Combat Engineer	2101	5.52	386	5.52	386	5.52
13B Cannon Crewman	4909	12.89	900	12.88	899	12.86
16S MANPADS Crewman	783	2.06	144	2.06	144	2.06
19E M60 Armor Crewman	580	1.52	106	1.52	107	1.53
19K M1 Armor Crewman	1822	4.78	335	4.79	334	4.78
27E TOW/Dragon Repairer	138	.36	25	.36	25	.36
29E Electronics Repairer	216	.57	40	.57	40	.57
31C Single Channel Radio Operator	970	2.55	178	2.55	179	2.56
51B Carpentry/Masonry Specialist	442	1.16	82	1.17	81	1.16
54E NBC Specialist	888	2.33	133	1.90	164	2.35
55B Ammunition Specialist	464	1.22	86	1.23	85	1.22
63B Light-Wheel Vehicle Mechanic	2121	5.57	390	5.58	391	5.59
67N Utility Helicopter Repairer	329	.86	60	.86	61	.87
71L Administrative Specialist	1944	5.10	358	5.12	357	5.11
76Y Unit Supply Specialist	2512	6.60	463	6.63	463	6.62
88M Motor Transport Operator	1540	4.04	284	4.06	283	4.05
91A Medical Specialist	3972	10.43	730	10.45	730	10.44
94B Food Service Specialist	3324	8.73	611	8.74	612	8.76
95B Military Police	4125	10.83	758	10.85	758	10.84
96B Intelligence Analyst	304	.80	56	.80	56	.80
Total ^b	38081		6959		6990	

^a Sample with both computer and paper-and-pencil batteries.

^b MOS codes were missing for 41 soldiers in the Initial Sample and 10 soldiers in Sample 2.

Table 3.2

Longitudinal Validation: Comparison of Initial Sample and Sample 2 Demographics to the Complete Sample^a by Race and Gender

Gender	Sample	Race				Total ^b
		Black	Hispanic	White	Other	
Female	Complete Sample N	1683	125	2917	142	4867
	% of Complete Sample Females	(34.6)	(2.6)	(59.7)	(2.9)	
	Initial Sample N	310	25	534	29	898
	% of Initial Sample Females	(34.5)	(2.8)	(59.5)	(3.2)	
	Sample 2 N	309	23	537	26	895
	% of Sample 2 Females	(34.5)	(2.6)	(60.0)	(2.9)	
Male	Complete Sample N	7670	1155	23169	1143	33140
	% of Complete Sample Males	(23.1)	(3.5)	(69.9)	(3.4)	
	Initial Sample N	1408	215	4259	209	6091
	% of Initial Sample Males	(23.1)	(3.5)	(69.9)	(3.4)	
	Sample 2 N	1409	211	4261	208	6089
	% of Sample 2 Males	(23.1)	(3.5)	(70.0)	(3.4)	
Total	Complete Sample N	9353	1283	26086	1285	38007
	% of Complete Sample	(24.6)	(3.4)	(68.6)	(3.4)	
	Initial Sample N	1718	240	4793	238	6989
	% of Initial Sample	(24.6)	(3.4)	(68.6)	(3.4)	
	Sample 2 N	1718	234	4798	234	6984
	% of Sample 2	(24.6)	(3.4)	(68.7)	(3.4)	

^a Sample with both computer and paper-and-pencil batteries.

^b Race codes were missing for 74 soldiers in the Complete Sample.

preliminary work had started on omitting individuals for whom data were incomplete or confusing.

Recall that the total sample includes approximately 38,000 soldiers with complete predictor data and more than 11,000 soldiers, mostly 11B Infantrymen, who did not take computer-administered predictors. Our subsamples were drawn from the sample of those who had complete predictor data. As a result, the subsamples contain proportionally fewer 11Bs than does the total sample. We do not think this is a serious issue for the kinds of analyses we completed. The 11B soldiers who took all predictors were not selected in any systematic way from the total group, and the 11B subsamples constitute a very sizeable proportion of the samples.

SCORING AND FORMING COMPOSITES OF PAPER-AND-PENCIL PREDICTORS

In this section analyses of the paper-and-pencil spatial tests are described. We begin by providing an overview of the tests and their psychometric properties, within a historical context. Next, two analyses are reported: (a) an investigation of methods for detecting random responses and (b) a study of alternate methods of scoring the spatial tests. Later we describe gender and race differences on the tests. Finally, analyses and conclusions regarding test composites are discussed.

Descriptions of Tests and Test History

Over the course of Project A and the Career Force project, we have accumulated a great deal of information about the six paper-and-pencil spatial tests. (See descriptions in Appendix A.) The tests have now been used in three major data collections and have undergone refinements at various stages of the research. The history of the six paper-and-pencil spatial tests is reviewed in Table 3.3. (See Peterson, 1987, or Peterson et al., 1990, for more complete descriptions of the development of the tests and the constructs they measure.)

The only spatial test that underwent changes between Concurrent and Longitudinal Validation was the Assembling Objects Test, a measure of spatial visualization. Four items were added and three items were revised because in Concurrent Validation the test had somewhat of a ceiling effect, as shown in Table 3.4. The ceiling effect for Assembling Objects dropped from Concurrent to Longitudinal Validation. This decrease could reflect differences between the two samples as well as changes to the test. However, as shown in Table 3.4, there were no large differences between the longitudinal and concurrent samples in means and standard deviations of test scores for the five unchanged tests. On most tests, the longitudinal sample performed slightly better and with slightly greater variability than did the concurrent sample. On the whole, performances of the two samples look very similar. Thus, the reduction of the ceiling effect on Assembling Objects appears to result from the changes in the test, not from differences between the samples.

Table 3.5 summarizes coefficient alpha and test-retest reliabilities obtained in the three major data collections: the Pilot Trial Battery or field test, the Trial Battery or Concurrent Validation, and the Experimental Battery or Longitudinal Validation. All of the tests consistently yield acceptable

Table 3.3

Changes in Cognitive Paper-and-Pencil Measures From Pilot Trial Battery to Trial Battery (Concurrent) to Experimental Battery (Longitudinal)

Test Name	Construct	No. of Items	Testing Time	Changes From Pilot Trial Battery Version ^a	Changes From Trial Battery Version
Assembling Objects Test	Spatial Visualization: Rotation	36	18.0 min.	Eight items were deleted	Four items were added and the time limit was increased by 2 minutes. Three items were revised.
Object Rotation Test	Spatial Visualization: Rotation	90	7.5 min.	No changes	No changes
Maze Test	Spatial Visualization: Scanning	24	5.5 min.	No changes	No changes
Orientation Test	Spatial Orientation	24	10.0 min.	Test title was changed from Orientation 2. Instructions Aid was added.	No changes
Map Test	Spatial Orientation	20	12.0 min.	Test title was changed from Orientation 3.	No changes
Reasoning Test	Induction	30	12.0 min.	Test title was changed from Reasoning 1.	No changes

^aPaper-and-pencil measures dropped from the Pilot Trial Battery were Shapes, Path, Reasoning 2, and Orientation 1.

Table 3.4

Cognitive Paper-and-Pencil Measures Completion Rates and Ceiling and Floor Effects: Concurrent Validation and Initial Longitudinal Samples

Test	Completion Rates			Floor/Ceiling Effects		Test Score Data	
	No. of Items	Mean	Median	Floor ^a	Ceiling ^b	Mean	SD
Assembling Objects Test							
Concurrent Validation ^c	32	30.97	32	.28	.70	23.29	6.71
Longitudinal Validation ^d	36	33.74	36	.03	.26	23.55	7.15
Object Rotation Test							
Concurrent Validation	90	70.13	72	-.62	.55	62.38	19.06
Longitudinal Validation	90	66.09	67	-.73	.47	59.13	20.15
Maze Test							
Concurrent Validation	24	17.48	18	.49	.40	16.39	4.77
Longitudinal Validation	24	17.86	18	.57	.55	16.95	4.85
Orientation Test							
Concurrent Validation	24	23.45	24	-.99	-.10	11.02	6.18
Longitudinal Validation	24	23.21	24	-.80	.11	12.25	6.21
Map Test							
Concurrent Validation	20	17.88	20	-1.06	-.23	7.67	5.51
Longitudinal Validation	20	17.29	19	-.99	-.23	7.86	5.45
Reasoning Test							
Concurrent Validation	30	28.46	30	.31	.07	19.07	5.67
Longitudinal Validation	30	27.86	30	.21	.08	19.53	5.44

^aFloor Effect = [(Mean Score - 2SD) - Chance] / SD, where chance is calculated using the median completion rate x p (item).

^bCeiling Effect = [(Mean Score + 2SD) - Maximum Possible] / SD.

^cN = 9,332-9,345.

^dN = 6,941-6,950, initial sample.

Table 3.5

Cognitive Paper-and-Pencil Measures: Reliability Comparisons Between Pilot Trial Battery, Trial Battery, and Experimental Battery Administrations

Test	Internal Consistency (Alpha)			Test-Retest	
	PTB ^a	TB ^b	EB ^c	PTB ^d	TB ^e
Assembling Objects ^f	.92	.90	.88	.74	.70
Object Rotation ^g	.97	.97	.98	.75	.72
Maze ^g	.89	.89	.90	.71	.70
Orientation	.88	.89	.89	.80	.70
Map	.90	.89	.88	.84	.78
Reasoning	.83	.86	.85	.64	.65

^aFort Knox sample, N = 290.

^bConcurrent sample, N = 9332-9345.

^cInitial longitudinal sample, N = 6754-6950.

^dN = 97-125.

^eN = 499-502.

^fContained 40 items in the Fort Knox field test and 32 items in the CV administration. Time limits were 16 minutes for both the PTB and TB. The EB contains 36 items and has an 18-minute time limit.

^gObject Rotation and Maze tests are designed to be speeded tests. Alpha is not an appropriate reliability coefficient but is reported here for consistency. Correlations between separately timed halves for the Pilot Trial Battery were .75 for Object Rotation and .64 for Maze (unadjusted).

reliability estimates. Also, similar levels of reliability were obtained across the different samples.

Test intercorrelations from the Concurrent and Longitudinal Validations are compared in Table 3.6. These correlations are similar in magnitude across the two samples (differing on the average $\pm .02$). Further examination suggests that the patterns of correlations are also similar across the two samples. For example, the Reasoning Test is most correlated with Assembling Objects and least correlated with Object Rotation in both samples.

These data lead us to two major conclusions. First, the six paper-and-pencil tests have consistently yielded acceptable test-retest and internal consistency reliability estimates. Second, the concurrent and longitudinal validation samples do not appear to differ substantially in terms of performance on these tests.

Evaluation of Methods for Screening Scores on the Six Spatial Tests

During Concurrent Validation, we examined two procedures designed to detect atypical response styles, the Personal Equation method and an Unlikely Response procedure. We decided not to use either procedure, mainly because we were unable to distinguish "might-be-random responders" from low-ability examinees. During Longitudinal Validation, we examined the usefulness of two different procedures: the Modified Caution Index (Harnisch & Linn, 1981) and a Runs Test.

The Modified Caution Index (MCI) is a measure of the extent to which a person responds in accordance with a Guttman model, or scalogram; it ranges from 0 to 1. Individuals whose response patterns do not fit the model (i.e., who get easy items wrong while getting difficult items right) receive high scores (near 1) on this index. Individuals whose response patterns fit the Guttman model (i.e., get easy items right and hard items wrong) obtain low (near 0) scores. This index is related to the Personal Equation, the correlation between item difficulty and item score (right/wrong), which we had decided against using with the CV data. Unfortunately, the Personal Equation is inherently related to total score (those who score either high or low tend to receive low scores on this index). The Modified Caution Index is reported to be less correlated with total score than other similar indexes (Harnisch & Linn, 1981).

A Runs Test is a count of the runs recorded by an individual respondent, where a run is a series of repeated item responses (e.g., 111111). Some careless responders will select the same response repeatedly, resulting in few runs over the items of the test. One problem with the Runs Test is that it is related to the number of items attempted by the examinee.

Procedures and Results

One major problem with applying the Modified Caution Index is that individuals who respond correctly to almost all of the items but miss one or two easier items do not fit the Guttman model and receive high scores; this suggests caution in interpreting their data, even though it is unlikely that they could have answered most items correctly if they were responding randomly.

Table 3.6**Cognitive Paper-and-Pencil Measures: Comparison of Correlations of Number Correct Score in Concurrent and Longitudinal Validations**

Concurrent Validation ^a Number Correct Score Intercorrelations					
Test	Object Rotation	Maze	Orientation	Map	Reasoning
Assembling Objects	.41	.51	.46	.50	.56
Object Rotation		.50	.37	.39	.38
Maze			.40	.44	.45
Orientation				.53	.48
Map					.52

^a N = 9332-9345.

Longitudinal Validation, Initial Sample ^b Number Correct Score Intercorrelations					
Test	Object Rotation	Maze	Orientation	Map	Reasoning
Assembling Objects	.46	.51	.50	.52	.56
Object Rotation		.51	.42	.42	.44
Maze			.41	.42	.48
Orientation				.54	.49
Map					.51

^b N = 6941-6950.

Consequently, the MCI must be interpreted in conjunction with information about the examinee's test score; an absolute cut score on the MCI alone would flag high-scoring individuals who probably did not respond randomly. Therefore, an accuracy score was computed for each individual on each of the six paper-and-pencil spatial tests to use in conjunction with the MCI and Runs Test. Accuracy is defined as the number correct divided by the number of items attempted; it ranges from 0 to 1. In itself, a high accuracy score indicates carefulness.

The accuracy score was used to define a range wherein scores appeared suspect on each test. One important assumption underlying our definition of this suspect score range is that poor test performance reflects low ability (and superior performance reflects high ability); performance well below that expected by chance is likely to be the result of low ability (just as performance well above chance is assumed to reflect ability). In sum, scores are not suspect simply because they are low. Individuals responding randomly to a test are likely to score within a range around the chance level of performance. We, therefore, defined the Suspect Score Range as chance-level test performance (expressed in the accuracy metric ± 2 Standard Error of Measurement on the test). The suspect score ranges and descriptive statistics of the accuracy scores on each test are provided in Table 3.7.

An MCI was computed for each individual on each of the six paper-and-pencil spatial tests. Descriptive statistics for these scores are provided in Table 3.8. We also prepared scatterplots showing MCI values plotted against accuracy scores. Frequency distributions of the MCI for individuals in the suspect score range on accuracy were examined to identify possible "break points" for an MCI cut score. No clear break was apparent. Therefore, we applied an absolute cut score on the MCI (i.e., $MCI > .50$) and examined the scores of individuals within the suspect score range on accuracy for each test. Numbers of individuals flagged by this "accuracy by MCI" rule are provided in Table 3.9. About 10 percent of the sample were flagged on one or more tests; most were flagged on just one. Shown in Table 3.10 are the mean AFQT scores for the total sample and for individuals flagged by the criteria. The AFQT scores of flagged cases were low in comparison to those of the total sample, indicating that low-ability examinees are much more prevalent in the group identified as possible random responders. Individuals flagged by the MCI criterion may, therefore, be low in ability and not necessarily random responders.

A Runs count was computed for each individual on each of the six paper-and-pencil spatial tests. Descriptive statistics for these scores are provided in Table 3.11. We prepared scatterplots of runs counts against accuracy scores, and examined the scores of individuals who had fewer than two runs (that is, marked the same response for every item) and scored within the suspect score range on accuracy for each test. As shown in Table 3.12, very few examinees were identified by these criteria.

Conclusions

We do not recommend using either the MCI or the Runs Test to screen longitudinal sample data. With regard to the MCI, the information (AFQT scores and spatial test scores) suggests that these individuals may be low in ability and not necessarily random responders. The runs counts did appear to identify

Table 3.7

Accuracy Score Statistics Used to Define Suspect Score Ranges on Six Paper-and-Pencil Tests

Test	N	Accuracy ^a			SEM ^c	Accuracy Suspect Score Range ^d	
		Mean	SD	Chance Level ^b		Min.	Max.
Assembling Objects	6941	.70	.18	.25	.06	.13	.37
Object Rotation	6950	.88	.15	.50	.02	.46	.54
Maze	6941	.95	.10	.25	.03	.19	.31
Orientation	6947	.52	.26	.20	.09	.02	.38
Map	6944	.44	.28	.13	.10	.00	.33
Reasoning	6948	.70	.18	.25	.07	.11	.39

^aAccuracy = Number Correct / Number Answered.

^bChance = 1.00 / Number of Response Alternatives per Item.

^cStandard Error of Measurement = SD $1 - r_x$, where r is the internal consistency reliability estimate computed on the Initial Sample data.

^dChance Level $\pm 2(\text{SEM})$.

Table 3.8

Longitudinal Validation Modified Caution Index: Descriptive Statistics on Six Paper-and-Pencil Tests for Initial Sample

Test	N	Modified Caution Index			
		Mean	SD	Median	IR ^a
Assembling Objects	6941	.29	.13	.28	.17
Object Rotation	6950	.14	.21	.05	.16
Maze	6941	.10	.21	.00	.09
Orientation	6947	.28	.15	.27	.21
Map	6944	.31	.19	.30	.25
Reasoning	6948	.20	.11	.18	.15

^aIR = Interquartile Range.

Table 3.9

Subjects Flagged by Applying Accuracy by Modified Caution Index Criteria on Six Paper-and-Pencil Tests: Initial Longitudinal Sample

Test	Initial Sample N	Total Number of Examinees Flagged on MCI by Accuracy Criteria ^a	Number of Examinees Flagged by Accuracy by MCI Criteria for One or More Tests			
			1	2	3	4
Assembling Objects	6941	45	31	11	2	1
Object Rotation	6950	15	10	5	0	0
Maze	6941	2	0	1	1	0
Orientation	6947	199	158	35	5	1
Map	6944	455	403	47	4	1
Reasoning	6948	44	27	13	3	1
Total N	6950	691	629	56	5	1

^aAccuracy within 2 SEM of chance and MCI greater than .50.

Table 3.10

AFQT Means for the Initial Longitudinal Sample and Examinees Flagged by the Modified Caution Index by Accuracy Criteria

	N	AFQT Score	
		Mean	SD
Initial Sample	6950	55.6	19.64
Examinees Flagged on Any Spatial Test	691	42.5	15.07
Examinees Flagged on One Test Only	629	42.8	15.08
Examinees Flagged on Two Tests	56	40.6	14.89
Examinees Flagged on Three Tests	5	27.0	4.47
Examinees Flagged on Four Tests	1	49.0	--

Table 3.11

Longitudinal Validation Runs Test: Descriptive Statistics on Six Paper-and-Pencil Tests for Initial Sample

Test	Initial Sample N	Runs Test			
		Mean	SD	Median	IR ^a
Assembling Objects	6941	27.8	2.46	28.0	2.82
Object Rotation	6950	53.6	7.53	55.4	4.38
Maze	6941	21.7	1.26	22.0	.93
Orientation	6947	20.3	2.11	21.0	3.00
Map	6944	17.7	1.98	18.0	2.33
Reasoning	6948	24.6	1.99	24.8	2.92

^aIR = Interquartile Range.

Table 3.12

Subjects Flagged by Applying Accuracy by Runs Test Criteria on Six Paper-and-Pencil Tests: Initial Longitudinal Sample

Test	Initial Sample N	Total Number of Examinees Flagged on Runs by Accuracy Criteria ^a	Number of Examinees Flagged on Runs by Accuracy Criteria for One or More Tests		
			Number of Tests Flagged 1	2	3
Assembling Objects	6941	2	1	0	1
Object Rotation	6950	6	6	0	0
Maze	6941	2	0	1	1
Orientation	6947	2	0	1	1
Map	6944	0	0	0	0
Reasoning	6948	0	0	0	0
Total N	6950	9	7	1	1

^aAccuracy within 2 SEM of chance; Runs, fewer than two runs.

some individuals who might have responded randomly, but so few were identified (nine out of about 6,950) that its use as a screen would have virtually no impact on the data.

Analysis of Alternative Scores

Another analyses goal was to determine the best way of scoring the spatial tests. During CV, two scores were examined, number correct and number wrong. In analyzing LV data both of these scores and an "accuracy" score were included. As noted earlier, accuracy is the number correct divided by number attempted; it ranges from 0 to 1. For some analyses, we also considered a speed score (number attempted/number possible). Accuracy and speed scores defined in this manner are related to proportion correct as follows:

$$\begin{array}{ccc} \text{Accuracy} & & \text{Speed} & & \text{Proportion Correct} \\ \frac{\text{Number Correct}}{\text{Number Attempted}} & & \frac{\text{Number Attempted}}{\text{Number Possible}} & - & \frac{\text{Number Correct}}{\text{Number Possible}} \end{array}$$

The number correct score is a measure of total test performance, taking both speed and accuracy into account. Both accuracy and "number wrong" are intended to assess the carefulness with which the individual completes the test, without regard to speed. But the two indexes are not equivalent, and it could be argued conceptually that accuracy is a purer measure of carefulness, although no research has compared the two psychometrically.² Analyses described below include both scores in order to further examine their psychometric properties and to facilitate comparison with CV data (for which number wrong scores were computed).

Analysis Procedures

Means and standard deviations of number correct, number wrong, number answered, speed, and accuracy scores for the LV Initial Sample were computed. As shown in Table 3.13, four of the six tests are primarily power tests (i.e., individuals achieve high speed scores). Object Rotation and Maze tests are the most speeded of the six cognitive paper-and-pencil measures (i.e., individuals achieve lower speed scores on these tests than the other tests). The high accuracy value for Maze Test, coupled with the speededness value, indicates that Maze Test is nearly a pure speed test: individuals tend to get all items reached correct. Moreover, the accuracy score is probably not a very good measure for tests that rely heavily on speed, like the Maze Test, since the accuracy mean is quite high and the variance is small.

²Accuracy may be a purer measure of carefulness than number wrong since the magnitude of the number wrong score (i.e., number attempted - number correct) is influenced by the number of items attempted. Consider two individuals who both have a 90 percent accuracy rate. One attempted 100 items and received a number wrong score of 10. If the other attempted 50 items, he or she would receive a number wrong score of 5. Although the two individuals would receive the same accuracy score (.90), they would not receive the same number wrong score.

Table 3.13

Longitudinal Validation: Five Alternative Scores on Cognitive Paper-and-Pencil Measures for Initial Sample

Test	No. of Items	Number Correct		Number Wrong		Number Answered		Speed ^a		Accuracy ^b	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Assembling Objects	36	23.55	7.15	10.20	6.37	33.74	4.93	.94	.14	.70	.18
Object Rotation	90	59.13	20.15	6.96	9.12	66.10	18.12	.73	.20	.88	.15
Maze	24	16.95	4.85	.91	1.89	17.86	4.73	.74	.20	.95	.10
Orientation	24	12.25	6.21	10.97	5.97	23.21	2.41	.97	.10	.52	.26
Map	20	7.86	5.45	9.43	5.14	17.29	3.60	.86	.18	.44	.28
Reasoning	30	19.53	5.44	8.33	5.00	27.06	3.74	.93	.13	.70	.18

Note: N = 6941-6950.

^aSpeed = Number Answered/Number of Items on the test. High values indicate less test speededness.^bAccuracy = Number Correct / Number Answered.

Table 3.14

Longitudinal Validation: Squared Multiple Regression Coefficients^a, Reliability Estimates, and Uniqueness Estimates for Cognitive Paper-and-Pencil Measures for Initial Sample

Test	R ²	Test-Retest Reliability ^b	Uniqueness ^c (Test-Retest)	Split-Half Reliability ^d	Uniqueness ^e (Split-Half)
Assembling Objects					
Number Correct	.28	.70	.42	.90	.62
Number Wrong	.22	.72	.50	.87	.65
Accuracy	.26	.73	.47	.86	.60
Object Rotation					
Number Correct	.20	.72	.52	.99	.79
Number Wrong	.12	.67	.55	.95	.83
Accuracy	.16	.69	.53	.94	.78
Maze					
Number Correct	.23	.70	.47	.96	.73
Number Wrong	.03	.43	.40	.81	.78
Accuracy	.05	.48	.43	.72	.67
Orientation					
Number Correct	.32	.70	.38	.89	.57
Number Wrong	.29	.69	.40	.88	.59
Accuracy	.32	.70	.38	.88	.56
Map					
Number Correct	.44	.78	.34	.90	.46
Number Wrong	.30	.66	.36	.88	.58
Accuracy	.42	.74	.32	.86	.44
Reasoning					
Number Correct	.32	.65	.33	.87	.55
Number Wrong	.26	.64	.38	.84	.58
Accuracy	.31	.66	.35	.84	.53

Note. N = 6857.

^aVersus all ASVAB subtests.

^bTest-retest reliabilities were computed using Concurrent Validation sample (N = 499-502).

^cUniqueness = Test-Retest Reliability - R².

^dSplit-half reliability estimates were computed using the odd/even method with the Spearman-Brown correction for test length.

^eUniqueness = Split-Half Reliability - R².

Squared multiple regression coefficients for paper-and-pencil scores with 10 ASVAB subtests, reliability estimates, and uniqueness estimates for the LV Initial Sample were also computed (see Table 3.14). The squared multiple regression coefficients are highest for number correct scores, ranging from .20 (Object Rotation Test) to .44 (Map Test). In general, the test-retest reliability estimates are higher for number correct scores than for number wrong and accuracy scores. The difference is largest for Maze Test, which has little variance in accuracy scores. The split-half reliability estimates are highest for number correct scores and lowest for accuracy scores. Again, the largest difference occurs for Maze Test, where the number wrong and accuracy scores were considerably less reliable than the total score.

Conclusions

The number wrong and accuracy scores might be useful measures on power tests, where their psychometric properties are similar to those for number correct, but they are not very useful for speeded tests. We recommend using the number correct score: (a) for the sake of consistency in scoring across all tests, and (b) because, overall, the number correct score has better psychometric properties.

Comparison of Gender and Race Subgroup Scores

To investigate gender and race differences on the tests, mean number correct scores and effect sizes by gender and by race were computed.

Gender Differences

Means, standard deviations, and effect sizes for gender subgroups from the CV sample and both LV samples (Initial and 2) are shown in Table 3.15. In the CV and LV samples, number correct scores were higher for men than for women on four of the six tests (Object Rotation, Maze, Orientation, and Map). Effect sizes for these four tests ranged from .21 to .38. Effect sizes for Assembling Objects Test were small. Reasoning Test scores favored women in the CV sample, but the effect size was essentially zero in the two LV samples.

Race Differences

Means, standard deviations, and effect sizes for race subgroups are shown in Table 3.16. In all three samples, scores were highest for whites and lowest for blacks. The Map Test and the Maze Test produced the largest differences; the Object Rotation Test produced the smallest differences. For Hispanics, the Map Test and Reasoning Test produced the largest differences while the Maze Test and Assembling Objects Test resulted in the smallest differences.

Table 3.15

Cognitive Paper-and-Pencil Measures: Means and Effect Sizes for Number Correct by Gender (CV Sample, LV Initial Sample, and LV Sample 2)

Test	Male			Female			Effect Size (d) ^a
	N	Mean	SD	N	Mean	SD	
Assembling Objects							
Concurrent Validation	8360	23.31	6.73	887	23.44	6.33	-.02
Longitudinal Validation Initial Sample	6049	23.60	7.18	892	23.19	6.90	-.06
Longitudinal Validation Sample 2	6057	23.61	7.15	888	23.03	7.12	.08
Object Rotation							
Concurrent Validation	8361	62.97	19.04	887	57.13	18.39	.31
Longitudinal Validation Initial Sample	6058	59.68	20.07	892	55.42	20.25	.21
Longitudinal Validation Sample 2	6058	59.78	19.88	888	54.19	19.23	.28
Maze							
Concurrent Validation	8362	16.55	4.74	885	15.04	4.81	.32
Longitudinal Validation Initial Sample	6050	17.17	4.82	891	15.48	4.86	.35
Longitudinal Validation Sample 2	6056	17.26	4.74	888	15.43	5.07	.38
Orientation							
Concurrent Validation	8358	11.17	6.24	887	9.75	5.35	.23
Longitudinal Validation Initial Sample	6055	12.52	6.27	892	10.39	5.40	.35
Longitudinal Validation Sample 2	6055	12.48	6.24	888	10.46	5.53	.33
Map							
Concurrent Validation	8362	7.84	5.56	885	6.32	4.90	.28
Longitudinal Validation Initial Sample	6052	8.07	5.49	892	6.45	4.94	.30
Longitudinal Validation Sample 2	6056	8.02	5.47	888	6.43	4.92	.29
Reasoning							
Concurrent Validation	8351	19.01	5.74	885	19.95	4.76	-.17
Longitudinal Validation Initial Sample	6056	19.52	5.48	892	19.59	5.16	-.01
Longitudinal Validation Sample 2	6059	19.58	5.45	888	19.57	5.04	.00

^a d is the standardized mean difference between males' and females' scores. A positive value indicates superior performance by males; a negative value indicates superior performance by females.

$$d = \frac{\bar{X}_m - \bar{X}_f}{S_p} \quad \text{where } S_p = \sqrt{\frac{(N_m - 1)S_m^2 + (N_f - 1)S_f^2}{(N_m - 1) + (N_f - 1)}}$$

Table 3.16

Cognitive Paper-and-Pencil Measures: Means and Effect Sizes for Number Correct by Race
(CV Sample, LV Initial Sample, and LV Sample 2)

Test	White				Black				Hispanic				Other			
	N	Mean	SD		N	Mean	SD		N	Mean	SD		N	Mean	SD	
				Effect Size ^a (d)				Effect Size ^a (d)				Effect Size ^a (d)				Effect Size ^a (d)
Assembling Objects																
Concurrent Validation	6007	24.75	6.04		2551	19.79	6.99	.78	334	23.87	6.19	.15	355	23.97	6.27	.13
Longitudinal Initial	4747	25.02	6.58		1710	19.39	7.16	.84	237	23.47	6.65	.24	236	24.03	6.79	.15
Longitudinal Sample 2	4749	24.99	6.52		1713	19.39	7.23	.83	233	23.35	7.03	.25	234	24.43	6.80	.09
Object Rotation																
Concurrent Validation	6006	66.48	17.50		2553	53.65	19.06	.71	334	59.34	20.69	.40	355	59.54	20.43	.39
Longitudinal Initial	4756	62.80	18.88		1710	49.45	20.11	.69	237	56.69	20.79	.32	236	57.39	20.65	.29
Longitudinal Sample 2	4750	62.32	18.86		1713	50.78	20.06	.60	233	56.39	19.59	.31	234	56.57	20.78	.30
Maze																
Concurrent Validation	6007	17.64	4.31		2551	13.36	4.50	.98	335	16.92	4.03	.17	354	16.97	4.81	.15
Longitudinal Initial	4749	18.06	4.42		1709	13.81	4.72	.94	237	17.04	4.47	.23	235	17.21	4.73	.19
Longitudinal Sample 2	4748	18.07	4.35		1713	14.04	4.87	.90	233	17.13	4.50	.22	234	17.49	4.62	.13
Orientation																
Concurrent Validation	6003	12.52	6.28		2553	7.79	4.46	.82	334	9.72	5.73	.45	355	10.45	6.14	.33
Longitudinal Initial	4752	13.65	6.09		1710	8.38	4.79	.91	237	11.51	5.80	.35	237	12.73	6.06	.15
Longitudinal Sample 2	4747	13.45	6.09		1713	8.96	5.10	.77	233	11.87	6.45	.26	234	11.59	6.31	.30
Map																
Concurrent Validation	6007	9.32	5.53		2550	4.16	3.47	1.03	335	6.32	4.96	.55	355	7.01	5.40	.42
Longitudinal Initial	4748	9.37	5.38		1711	4.03	3.47	1.08	237	6.03	4.66	.62	237	7.11	5.00	.42
Longitudinal Sample 2	4748	9.16	5.41		1714	4.26	3.64	.98	233	7.16	5.25	.37	233	7.21	5.17	.36
Reasoning																
Concurrent Validation	6002	20.41	5.15		2547	16.34	5.64	.77	333	17.39	6.06	.58	354	18.26	5.99	.41
Longitudinal Initial	4754	20.62	5.02		1710	16.63	5.50	.77	237	18.49	5.46	.42	236	19.58	5.32	.21
Longitudinal Sample 2	4750	20.67	4.92		1714	16.79	5.54	.76	233	18.13	6.09	.51	234	19.32	5.31	.27

^a d is the standardized mean difference between two subgroups' scores. All effect sizes in this table are relative to the white subgroup.

For example, a positive effect size for blacks indicates superior performance by whites, and a negative value indicates superior performance by blacks.

$$d = \frac{\bar{X}_W - \bar{X}_B}{S_p} \quad \text{where } S_p = \sqrt{\frac{(N_W - 1)S_W^2 + (N_B - 1)S_B^2}{(N_W - 1) + (N_B - 1)}}$$

Analyses and Conclusions Regarding Composite Formation

Principal factor analyses (using squared multiple correlations as the communality estimates) were conducted on the six cognitive paper-and-pencil number correct scores, and the eigenvalues were compared to parallel analysis estimates of eigenvalues for random data (Allen & Hubbard, 1986; Humphreys & Montanelli, 1975; Montanelli & Humphreys, 1976). The parallel analysis suggested that one or at most two factors should be retained for the spatial tests. Table 3.17 presents the two-factor orthogonal varimax rotation solution for both the longitudinal Initial sample and the concurrent sample. The two solutions are highly similar.

In this exploratory analysis, Object Rotation Test and Maze Test (speeded tests) load on the second factor and all other tests load on the first. The second factor appears to be a method (speededness) factor that does not reflect a meaningful homogeneous construct.

When factored with other cognitive measures (i.e., the ASVAB subtests and/or computer measures), the spatial tests consistently form a single factor of their own. Tables 3.18 and 3.19 report factor analyses of spatial tests and ASVAB subtests for the concurrent and longitudinal samples respectively.

Dr. Lloyd Humphreys, a member of the project's Scientific Advisory Group, pointed out other aspects of data (regarding the gender differences on the spatial tests) suggesting that more than one factor might be reasonable (personal communication, March 1990). Specifically, he noted that there is little or no gender difference on the Reasoning Test and Assembling Objects Test. Gender differences on the other tests are consistent with those found in spatial abilities research (about one-quarter to one-third of a standard deviation). This could suggest grouping the two gender-neutral tests to form a composite. We conducted confirmatory and second-order analyses to further investigate this matter.

Analysis Procedures

Using LISREL (Joreskog & Sorbom, 1986) four models were compared. Model 1 had one composite formed by all six spatial tests. Model 2 included two composites: (a) visualization speed (Maze Test, Object Rotation Test) and (b) power (all other tests). The third model also had two composites: (a) figural reasoning (Assembling Objects Test, Reasoning Test) and (b) general spatial (all other tests). The fourth model had three composites: (a) visualization speed (Maze Test, Object Rotation Test), (b) figural reasoning (Assembling Objects Test, Reasoning Test), and (c) orientation (Orientation Test, Map Test).

The LISREL analyses, as shown in Table 3.20, suggested that the second or fourth models might be useful ways to summarize these scores. We expected model 2 to fit well because it is the model suggested by exploratory analyses (i.e., the power/speed distinction). For model 4, the chi-square value was reduced substantially (from 235.62 to 19.62) with a loss of three degrees of freedom. The correlations between the factors for model 4 (see the Phi matrix) are not extremely high and suggest that the three factors may measure somewhat different constructs.

Table 3.17

Cognitive Paper-and-Pencil Measures: Factor Loadings for Number Correct, Principal Factor Analyses^a Two-Factor Solution (CV Sample and LV Initial Sample)

Test	<u>Concurrent Validation</u>		h^2 ^b
	Factor I	Factor II	
Map	.60	.37	.50
Reasoning	.59	.40	.50
Orientation	.56	.34	.43
Assembling Objects	.54	.47	.51
Maze	.38	.57	.48
Object Rotation	.32	.52	.38
Eigenvalue	1.56	1.24	2.80

Note. N = 7939.

Test	<u>Longitudinal Validation Initial Sample</u>		h^2 ^b
	Factor I	Factor II	
Map	.59	.38	.49
Orientation	.57	.37	.46
Assembling Objects	.55	.49	.54
Reasoning	.54	.46	.50
Maze	.38	.57	.47
Object Rotation	.36	.54	.42
Eigenvalue	1.54	1.35	2.88

Note. N = 6929.

^a Squared multiple correlations were used as initial communality estimates. Orthogonal varimax rotation.

^b h^2 = communality (sum of squared factor loadings) for variables.

Table 3.18

Concurrent Validation Cognitive Paper-and-Pencil Measures: Factor Loadings for Number Correct and ASVAB Subtests, Principal Factor Analysis^a Five-Factor Solution

Test	Factor I Spatial	Factor II General Ability	Factor III Technical	Factor IV Numerical	Factor V Speed	h^2 ^b
Assembling Objects	.68	.15	.09	.06	.16	.52
Maze	.66	.07	.15	.18	.00	.56
Reasoning	.61	.28	.06	.10	.26	.49
Object Rotation	.58	.05	.19	.12	.00	.38
Orientation	.56	.21	.17	.01	.23	.44
Map	.56	.33	.23	.10	.29	.53
ASVAB						
Word Knowledge	.15	.75	.21	-.01	.13	.65
General Science	.25	.67	.36	.00	.16	.67
Paragraph Comprehension	.14	.64	.18	.10	.12	.49
Auto/Shop	.26	.31	.61	-.13	.05	.56
Electronics Information	.20	.36	.59	-.05	.12	.54
Mechanical Comprehension	.46	.33	.53	-.05	.24	.66
Number Operations	.06	-.03	-.07	.64	.11	.44
Coding Speed	.16	.08	-.02	.60	.03	.39
Mathematics Knowledge	.31	.38	.16	.30	.54	.65
Arithmetic Reasoning	.35	.36	.23	.24	.52	.64
Eigenvalue	2.89	2.31	1.45	1.03	0.94	8.61

Note. N = 7884.

^aInitial communality estimates = squared multiple correlations.
Orthogonal varimax rotation.

^b h^2 = communality (sum of squared factor loadings) for variables.

Table 3.19

Longitudinal Validation Cognitive Paper-and-Pencil Measures: Factor Loadings for Number Correct and ASVAB Subtests, Principal Factor Analysis^a Five-Factor Solution (Initial Sample)

Test	Factor I Spatial	Factor II General Ability	Factor III Technical	Factor IV Numerical	Factor V Speed	h^2 ^b
Assembling Objects	.68	.18	.10	.15	.04	.53
Maze	.65	.10	.14	.02	.18	.48
Reasoning	.64	.23	.03	.26	.09	.54
Object Rotation	.62	.08	.16	.03	.14	.44
Orientation	.59	.18	.18	.27	-.04	.49
Map	.56	.33	.19	.31	.06	.56
ASVAB						
Mechanical Comprehension	.47	.38	.47	.29	-.07	.68
Arithmetic Reasoning	.34	.33	.21	.58	.15	.63
Mathematics Knowledge	.30	.37	.07	.60	.22	.64
Auto/Shop	.28	.31	.61	.07	-.18	.58
Electronics Information	.23	.48	.55	.15	-.09	.62
General Science	.21	.73	.28	.20	-.09	.70
Word Knowledge	.17	.75	.17	.13	-.08	.64
Paragraph Comprehension	.17	.60	.11	.15	.12	.44
Coding Speed	.14	.04	-.05	.00	.62	.41
Number Operations	.06	-.08	-.07	.15	.66	.47
Eigenvalue	3.02	2.42	1.20	1.17	1.03	8.85

Note. N = 6857.

^a Initial communality estimates = squared multiple correlations.
Orthogonal varimax rotation.

^b h^2 = communality (sum of squared factor loadings) for variables.

Table 3.20

LISREL Runs on Initial Longitudinal Sample Spatial Test Data to Examine Four Alternate Composite Models^a

Input Correlation Matrix (N = 4723)

Test	Assembling Objects	Map	Maze	Object Rotation	Orientation	Reasoning
Assembling Objects	1.00					
Map	.51	1.00				
Maze	.48	.40	1.00			
Object Rotation	.44	.40	.50	1.00		
Orientation	.49	.52	.39	.40	1.00	
Reasoning	.55	.50	.45	.41	.47	1.00

Definitions of Alternate Composite Models

- 1) One composite of all spatial tests.
- 2) Two composites: a) speed (Maze Test, Object Rotation Test) and b) power (all other tests).
- 3) Two composites: a) figural (Assembling Objects, Reasoning Test) and b) spatial (all other tests).
- 4) Three composites: a) speed (Maze Test, Object Rotation Test), b) figural (Assembling Objects, Reasoning Test), c) orientation (Orientation Test, Map Test).

Generalized Least Squares LISREL Results

Model	Coefficient of Determination	df	Chi Square	Goodness of Fit	Adjusted Goodness of Fit	Root Mean Square Residual
1	.869	9	235.62	.983	.961	.033
2	.902	8	79.57	.994	.985	.017
3	.884	8	217.89	.985	.960	.033
4	.917	6	19.62	.999	.995	.008

Phi Matrixes for Models 2, 3, and 4: Estimates of True Score Correlations

Model 2			Model 3			Model 4		
Power	Speed		Figural	Spatial		Speed	Figural	Orientation
Power	1.0		Figural	1.0		Speed	1.0	
Speed	.85	1.0	Spatial	.96	1.0	Figural	.86	1.0
						Orientation	.78	.92
								1.0

^a May 1990.

Even though the LISREL analyses might support the use of three composites, all the exploratory work supports a one-factor solution (see Tables 3.17, 3.18, and 3.19). With this in mind, Lloyd Humphreys (personal communication, March 1990) suggested we do a "second-order analysis and use the Schmid-Leiman transformation to place both first order and second order in a single order consisting of a general factor and orthogonal group factors." We followed through with his suggestion. The results of the Schmid-Leiman transformation (Schmid & Leiman, 1957) are provided in Table 3.21. As shown, all tests have large loadings on the second-order general factor. Loadings on the Speed and Orientation specific factors are small, and loadings on the Figural factor are essentially zero, suggesting that virtually all reliable variance in the Assembling Objects and Reasoning tests is tapped by the general factor.

Results and Conclusions

The decision on the number of spatial test composites requires considering practical concerns as well as research findings. Some reasons for using more than one composite are (a) the Army may wish in the future to administer fewer than six of the paper-and-pencil tests; (b) it might be useful to have a gender-neutral spatial composite (Reasoning); and (c) the specific factors might predict different criteria.

There are also several reasons for using one composite for all tests. First, including more than one spatial composite in prediction equations will reduce degrees of freedom, a consideration that may be important for within-MOS analyses where *N*s may be small. Second, all six tests have "strong" loadings on the general factor (ranging from .62 for Maze to .75 for Assembling Objects); moreover, the loadings on the specific factors are moderate to small.

Third, the constructs defined by alternate solutions are not highly meaningful. The speeded tests might share variance because they are both measures of visualization or because of their speededness; we speculate that they are defining a speed (method) factor. The orientation factor does not emerge in exploratory single-order analyses, and the two tests loading on that factor have fairly small loadings (see Table 3.21). Also, it makes sense that the Reasoning Test represents a general reasoning construct, but we are unsure why Assembling Objects would also define this construct. Lloyd Humphreys (personal communication, 1990) suggests that broad factors are likely to be better predictors than narrow factors, though he also knows of instances where specific factors were very useful.

In the situation for which these tests are intended (i.e., selection/classification of applicants into entry-level enlisted Army occupations), we do not know of any reason to expect a spatial speed factor to be particularly useful. The figural factor is better measured by a unit-weighted composite of all six tests than by a composite of two tests. The orientation factor does not appear to explain much variance unique from the general spatial factor. Therefore, for purposes of validation analyses, we recommend forming one unit-weighted composite of the six spatial test scores.

Table 3.21

Second-Order Analysis of Spatial Test Scores: Schmid-Leiman Transformation

Input Correlation Matrix (N = 4723)

Test	Assembling Objects	Map	Maze	Object Rotation	Orientation	Reasoning
Assembling Objects	1.00					
Map	.51	1.00				
Maze	.48	.40	1.00			
Object Rotation	.44	.40	.50	1.00		
Orientation	.49	.52	.39	.40	1.00	
Reasoning	.55	.50	.45	.41	.47	1.00

Loadings on the three oblique first-order factors

Test	Speed	Figural	Orientation
Assembling Objects	.000	.756	.000
Map	.000	.000	.739
Maze	.724	.000	.000
Object Rotation	.686	.000	.000
Orientation	.000	.000	.708
Reasoning	.000	.723	.000

Correlations between oblique first-order factors

	Speed	Figural	Orientation
Speed	1.00		
Figural	.86	1.00	
Orientation	.78	.92	1.00

Loadings for first-order factors on the second-order factor

Speed	.862
Figural	.996
Orientation	.927

Results

Test	General Factor	Specific Factors		
		Speed	Figural	Orientation
Assembling Objects	.753	.000	.065	.000
Map	.685	.000	.000	.278
Maze	.624	.367	.000	.000
Object Rotation	.592	.347	.000	.000
Orientation	.656	.000	.000	.266
Reasoning	.720	.000	.062	.000

For applied settings in which the Army may wish to use fewer than six tests to test spatial abilities, we suggest using the Assembling Objects Test. It is a good measure of the general factor, and it consistently yields smaller gender and race differences than the other tests.

Data from Sample 2 were used as a replicate. Table 3.6 demonstrated that the correlations of the six tests are very similar across the concurrent and longitudinal samples and suggested that one, or at the most two, factors should be retained. The two-factor solution from Sample 2 is presented alongside the same solution from the Initial sample in Table 3.22. The factor structures from the two longitudinal samples are highly similar.

Table 3.22

Comparison of Cognitive Paper-and-Pencil Test Factor Loadings for Two Longitudinal Validation Samples^a

Test	Factor I Initial/Sample 2	Factor II Initial/Sample 2	h^2 ^b Initial/Sample 2
Map	.59/.58	.38/.38	.49/.48
Orientation	.57/.56	.37/.35	.46/.44
Assembling Objects	.55/.54	.49/.50	.54/.54
Reasoning	.54/.54	.46/.42	.50/.47
Maze	.38/.37	.57/.55	.47/.43
Object Rotation	.36/.34	.54/.52	.42/.38
Eigenvalue	1.54/1.49	1.35/1.26	2.88/2.75

Note. Initial sample N = 6929, Sample 2 N = 6436.

^a Principal factor analysis with varimax rotation.

^b h^2 = communality (sum of squared factor loadings) for variables.

SCORING AND FORMING COMPOSITES OF COMPUTER-ADMINISTERED PREDICTOR SCORES

This section summarizes several analyses of the computer-administered measures: (a) analysis of sources of within-subject variance in test scores and the implications for test scoring, (b) analysis of alternate methods of scoring, (c) data screening rules, (d) analysis of subgroup scores, and (e) the analyses directed toward forming composites of scores of the computer-administered predictors.

Test Descriptions

The ten computer-administered tests and the constructs they were designed to measure are listed in Table 3.23. Descriptions and sample items are in Appendix A. Also provided in Table 3.23 is a summary of changes made in the test battery from the field test to Longitudinal Validation. None of the computer tests changed substantively between Concurrent and Longitudinal Validation, although some minor changes were made in the instructions and the test software. (See Peterson, 1987, or Peterson, et al., 1990, for more complete descriptions of the tests and the constructs they measure.)

The full computer battery takes about one hour to administer. The mean test times for each test for the Concurrent and Initial Longitudinal Validation samples are shown in Table 3.24. The two samples are highly similar; new recruits (the longitudinal sample) took about the same amount of time as the first-tour incumbents (Concurrent Validation).

The ten tests can be divided into two groups: (a) cognitive/perceptual tests and (b) psychomotor tests. The cognitive/perceptual tests include: Simple Reaction Time, Choice Reaction Time, Perceptual Speed and Accuracy, Short-Term Memory, Target Identification, and Number Memory. With the exception of Number Memory, three scores are recorded for each item on these tests, decision time, movement time, and correct/incorrect; Number Memory has three time scores and a proportion correct score. The item scores on the psychomotor tests (Target Tracking 1, Target Tracking 2, Target Shoot, and Cannon Shoot) are in either distance or time units. In either case, the measurement reflects the precision with which the examinee has tracked or shot at the target.

Variance Analysis of Selected Computer Test Scores

The major objective of the variance analyses was to investigate the effects of parameters on test performance and the implications of these effects for test scoring. The items on most of the computer tests can be described in terms of several parameters. For example, the Perceptual Speed and Accuracy Test has three defining parameters: (a) the number of characters in the stimulus (three levels--two, five, or nine characters), (b) the type of stimulus (letters, numbers, symbols, or a mix of these), and (c) position of correct response, the position of the button that should be pressed to get the item correct, which we called probe status (two levels). One item might include a stimulus that has two characters that are symbols and a different comparison stimulus (e.g., &S compared to &*); in this case the examinee must press the "different" button to answer correctly. Another item may have nine characters that are mixed in type with an identical comparison stimulus (e.g., a*7//c5+n compared to a*7//c5+n). Figure 3.2 lists the item parameters relevant to each test.

During the field test and Concurrent Validation, we found that the test item parameters influence within-subject variance in decision time, sometimes to a large degree. For example, for Perceptual Speed and Accuracy (PS&A), the parameter "number of characters" contributes over 90 percent of the within-subject variance in decision time; decision time increases as the number of characters in the stimulus increases. These findings have important implications for test scoring; for example, should mean decision times for

Table 3.23

Longitudinal Validation: Changes in Computer-Administered Measures From Pilot Trial Battery to Trial Battery to Experimental Battery

Test Name	Construct	No. of Items	Time Limits	Changes for the Trial Battery (Concurrent)	Changes for the Experimental Battery (Longitudinal) ^a
Demographics					
Simple Reaction Time	Reaction Time	15 ^b	9 seconds per item	Eliminated typing, race, and age items. Retained SSN and video experience items.	None
Choice Reaction Time	Reaction Time	30	9 seconds per item	No changes.	None
Perceptual Speed & Accuracy	Perceptual Speed & Accuracy	36	9 seconds per item	Increased number of items from 15 to 30.	None
Target Identification	Perceptual Speed & Accuracy	36	9 seconds per item	Reduced items from 48 to 36. Eliminated word items.	None
Short Term Memory	Memory	36	9 seconds per item	Reduced items from 48 to 36. Eliminated moving items. Allowed stimuli to appear at more angles of rotation	None
Number Memory	Memory	28	12 seconds (operations) 9 seconds (final response)	Reduced items from 48 to 36. Established a single item presentation and probe delay period.	None
Target Tracking 1	Precision/Steadiness	18	None	Reduced length of items. Increased number of items from 27 to 28.	None
Target Shoot	Precision/Steadiness	30	None	Reduced number of items from 27 to 18. Increased item difficulty.	None
Target Tracking 2	Multilimb Coordination	18	None	Reduced number of items from 40 to 30 by eliminating extremely easy and extremely difficult items. (Also, categorized items by difficulty, crosshair speed, and target/crosshair ratio.)	None
Cannon Shoot	Movement Judgment	36	None	Reduced number of items from 27 to 18. Increased item difficulty.	None
				Reduced number of items from 48 to 36.	None

^aMinor wording changes in instructions were made to all tests. Minor item display changes were made to the software for all tests.

^bThe Simple Reaction Time Test was used as a familiarization test for the equipment; the first five items were for practice only, leaving 10 scorable items.

Table 3.24

Concurrent and Longitudinal Validations: Mean Time to Read Instructions and Complete Test Items for Computer-Administered Tests

Test	Period	Concurrent		Longitudinal (Initial Sample)	
		N	Mean	N	Mean
Demographics	Instruction	9273	0 min 19 sec	6999	0 min 19 sec
	Test Items		4 min 1 sec		4 min 3 sec
	Total		4 min 20 sec		4 min 22 sec
Simple Reaction Time	Instruction	9271	N/A	6996	1 min 38 sec
	Test Items		N/A		0 min 45 sec
	Total		2 min 20 sec		2 min 23 sec
Choice Reaction Time	Instruction	9272	1 min 0 sec	6993	1 min 11 sec
	Test Items		1 min 58 sec		1 min 57 sec
	Total		2 min 58 sec		3 min 8 sec
Perceptual Speed & Accuracy	Instruction	9266	1 min 44 sec	6990	1 min 41 sec
	Test Items		3 min 25 sec		3 min 22 sec
	Total		5 min 9 sec		5 min 3 sec
Target Identification	Instruction	9232	1 min 29 sec	6952	1 min 35 sec
	Test Items		2 min 34 sec		2 min 28 sec
	Total		4 min 3 sec		4 min 3 sec
Short Term Memory	Instruction	9269	2 min 41 sec	6994	2 min 41 sec
	Test Items		4 min 54 sec		4 min 51 sec
	Total		7 min 35 sec		7 min 32 sec
Number Memory	Instruction	9256	3 min 14 sec	6978	3 min 17 sec
	Test Items		6 min 18 sec		6 min 17 sec
	Total		9 min 32 sec		9 min 34 sec
Target Tracking 1	Instruction	9256	3 min 37 sec	6993	3 min 41 sec
	Test Items		3 min 47 sec		3 min 56 sec
	Total		7 min 24 sec		7 min 37 sec
Target Shoot	Instruction	9208	1 min 44 sec	6935	1 min 51 sec
	Test Items		3 min 6 sec		3 min 2 sec
	Total		4 min 50 sec		4 min 54 sec
Target Tracking 2	Instruction	9254	2 min 14 sec	6985	2 min 16 sec
	Test Items		3 min 47 sec		3 min 46 sec
	Total		6 min 1 sec		6 min 2 sec
Cannon Shoot	Instruction	9238	3 min 8 sec	6967	3 min 4 sec
	Test Items		3 min 43 sec		3 min 42 sec
	Total		6 min 51 sec		6 min 46 sec
Total Time for Battery	Total Test Time	9200	61 min 3 sec	6935	61 min 16 sec

Perceptual Speed and Accuracy

- Stimulus type (alpha, numeric, symbolic, mixed)
- Number of characters in each set of stimuli (2, 5, 9)
- Same vs different correct answer

Target Identification

- Difficulty (2 levels)

Short-Term Memory

- Stimulus type (alpha, symbolic)
- Number of characters in each set of stimuli (1, 3, 5)
- Probe status (in or not in)

Number Memory

- Operation by problem part (add, subtract, multiply, divide)

Target Tracking 1

- Crosshair speed
- Ratio of crosshair speed to target speed
- Number of turns in path

Target Shoot

- Item difficulty
- Crosshair speed
- Ratio of crosshair speed to target speed

Target Tracking 2

- Crosshair speed
- Ratio of crosshair speed to target speed
- Number of turns in path

Cannon Shoot

- Distance from target onset to optimal point of fire
- Distance from cannon to stimulus
- Angle of stimulus path

Note. Simple and Choice Reaction Time tests are not listed here because they are simple tests that do not have parameters with meaningful scoring implications.

Figure 3.2. Summary of computer-administered test parameters.

PS&A simply be computed across all items, or be computed as the mean of mean scores on groups of items defined by "number of characters." Thus, the variance analyses are not purely descriptive; they provide information about how the test might best be scored.

Procedure

As mentioned, analyses of variance had been computed during Concurrent Validation and the field tests; the Longitudinal Validation variance analyses were more confirmatory in nature than exploratory. To conserve resources, a random sample of 700 cases was drawn from the Initial Longitudinal sample. Variance components were computed and compared against analyses of the full CV sample (computed during CV). This comparison served as a vehicle for examining the robustness of results across sampling strategies.

Variance estimates were computed for all the tests except Simple and Choice Reaction Time Tests and the Number Memory Test. Recall that the goal of these analyses was to investigate the implications of parameter effects on test scoring; moreover, we conducted these analyses because we were considering scoring these tests in ways that differ from their CV scoring. Simple and Choice Reaction Time Tests are simple tests that do not have parameters with meaningful scoring implications. The Number Memory Test is scored by parameters in accordance with conceptual underpinnings that drove its development (Peterson, et al., 1987); it was not included in analyses because its scoring (parameter-wise) was not under reassessment.

The variances estimates appear in Tables 3.25-3.31, with the corresponding estimates computed for the concurrent sample. Table 3.32 shows Number Memory variance results obtained during Concurrent Validation.

Results

For all the analysis of variance tests, the CV and LV variance component estimates are highly similar, suggesting that the effect of parameters on within-subject scores is consistent across samples. On Perceptual Speed and Accuracy, for example, the number of characters in the stimulus explains about 94 percent of the within-subject variance in decision time for both concurrent and longitudinal samples (Table 3.25).

Parameters that account for a large portion of variance in decision time for the reaction time tests are the number of characters in each set of stimuli for Perceptual Speed and Accuracy and Short-Term Memory (see Tables 3.25 and 3.26). Parameters that account for relatively large portions of variance in the mean log (distance + 1) score are crosshair speed for Target Tracking 1 and Target Tracking 2 (see Tables 3.27 and 3.28) and, for Target Shoot (Table 3.29), item difficulty and crosshair speed. Cannon Shoot is a complex test; several parameters and interactions affect within-subject variance (Table 3.30).

Decision time on the Target Identification Test is influenced by an item difficulty parameter created on the basis of the visual similarities of response alternatives on this test (Table 3.31). Here, difficulty accounts

Table 3.25

Computer-Administered Tests: Analysis of Variance Due to Item Parameters^a -- Perceptual Speed and Accuracy

		Concurrent Validation			Longitudinal Validation		
Variance Components		Variance ^b	Prop. of W/In	Prop. of Total	Variance ^b	Prop. of W/In	Prop. of Total
Proportion Correct:	Total Variance	.119	--	1	.133	--	1
	Within Subject Variance	.022	1	.181	.025	1	.187
	A. Stimulus Type	.001	.035	.006	.001	.029	.005
	B. No. of Characters	.004	.173	.032	.004	.156	.029
	C. Probe Status	.011	.493	.090	.014	.549	.103
	A x B	-.001	-.050	-.009	-.001	-.054	-.010
	A x C	.002	.079	.014	.002	.094	.018
	B x C	.004	.182	.033	.004	.154	.029
	A x B x C	.002	.087	.016	.002	.072	.014
	Subject I.D.	.004	--	.034	.005	--	.035
	Error	.093	--	.784	.103	--	.778
Decision Time:	Total Variance	30253.64	--	1	30106.07	--	1
	Within Subject Variance	19669.18	1	.650	19361.38	1	.643
	A. Stimulus Type	-605.26	-.031	-.020	-655.36	-.034	-.022
	B. No. of Characters	18453.46	.938	.610	18155.54	.938	.603
	C. Probe Status	-856.10	-.044	-.028	-961.55	-.050	-.032
	A x B	984.13	.050	.033	1149.63	.059	.038
	A x C	1299.89	.066	.043	1469.64	.076	.049
	B x C	1973.11	.100	.065	2006.68	.104	.067
	A x B x C	-1580.05	-.080	-.052	-1803.29	-.093	-.060
	Subject I.D.	3453.89	--	.114	3472.88	--	.115
	Error	7130.57	--	.236	7271.87	--	.242
Movement Time:	Total Variance	679.61	--	1	717.57	--	1
	Within Subject Variance	17.81	1	.026	24.74	1	.034
	A. Stimulus Type	-.23	-.013	.000	-.16	-.007	-.001
	B. No. of Characters	.83	.047	.001	-.45	-.018	-.001
	C. Probe Status	15.46	.868	.023	20.75	.839	.029
	A x B	.88	.049	.001	1.97	.080	.003
	A x C	.39	.022	.001	.31	.013	.000
	B x C	.57	.032	.001	2.71	.110	-.004
	A x B x C	-.09	-.005	.000	-.42	-.017	-.001
	Subject I.D.	113.94	--	.168	95.87	--	.134
	Error	547.85	--	.806	596.97	--	.832

Note. CV N = 9423; LV N = 700. Missing data not imputed.

^aNegative values suggest that the model overexplains the total variance.

^bDecision Time and Movement Time in tenths of seconds.

Table 3.26

Computer-Administered Tests: Analysis of Variance Due to Item Parameters^a -- Short-Term Memory

		Concurrent Validation			Longitudinal Validation		
Variance Components		Variance ^b	Prop. of W/In	Prop. of Total	Variance ^b	Prop. of W/In	Prop. of Total
Proportion Correct:	A. Stimulus Type	.002	.207	.022	.002	.290	.024
	B. No. of Characters	.001	.091	.010	.001	.111	.009
	C. Probe Status	.001	.052	.005	.000	.037	.003
	A x B	.002	.182	.019	.001	.126	.010
	A x C	.001	.116	.012	.001	.100	.008
	B x C	.000	.020	.002	-.001	-.067	-.006
	A x B x C	.004	.332	.035	.003	.403	.033
	Subject I.D.	.004	--	.034	.006	--	.053
	Error	.089	--	.861	.089	--	.864
	Total Variance	.103	--	1	.103	--	1
Within Subject Variance		.011	1	.105	.009	1	.083
Decision Time:	A. Stimulus Type	9.58	.043	.006	26.04	.111	.012
	B. No. of Characters	145.75	.655	.094	148.43	.634	.071
	C. Probe Status	45.04	.202	.029	59.35	.253	.028
	A x B	18.18	.082	.012	9.66	.041	.005
	A x C	1.42	.006	.001	-4.94	-.021	-.002
	B x C	2.87	.013	.002	9.97	.043	.005
	A x B x C	-.36	-.002	.000	-14.27	-.061	-.007
	Subject I.D.	529.90	--	.342	510.40	--	.243
	Error	796.99	--	.514	1354.71	--	.645
	Total Variance	1549.35	--	1	2099.36	--	1
Within Subject Variance		222.48	1	.144	234.24	1	.112
Movement Time:	A. Stimulus Type	.77	.163	.001	1.42	.144	.002
	B. No. of Characters	-.52	-.111	.000	-.52	-.053	-.001
	C. Probe Status	2.80	.593	.004	5.92	.601	.007
	A x B	.03	.006	.000	.72	.073	.001
	A x C	.03	.006	.000	-.14	-.014	.000
	B x C	1.49	.315	.002	3.46	.351	.004
	A x B x C	.13	.027	.000	-1.01	-.102	-.001
	Subject I.D.	163.68	--	.261	152.09	--	.183
	Error	459.89	--	.732	670.79	--	.806
	Total Variance	628.29	--	1	832.66	--	1
Within Subject Variance		4.72	1	.008	9.86	1	.012

Note. CV N = 9423; LV N = 700. Missing data not imputed.

^aNegative values suggest that the model overexplains the total variance.

^bDecision Time and Movement Time in tenths of seconds.

Table 3.27

Computer-Administered Tests: Analysis of Variance Due to Item Parameters^a--
Target Tracking Test 1

Log Distance : + 1	Variance Components	Concurrent Validation			Longitudinal Validation		
		Variance	Prop. of W/In	Prop. of Total	Variance	Prop. of W/In	Prop. of Total
	Total Variance	.488	--	1	.449	--	1
	Within Subject Variance	.140	1	.286	.127	1	.283
	A. Crosshair Speed	.104	.748	.214	.102	.801	.227
	B. Speed Ratio	.028	.197	.056	.020	.155	.044
	C. Number of Turns	.002	.016	.005	.002	.012	.003
	A x B	.003	.021	.006	.003	.022	.006
	A x C	.001	.005	.001	.000	-.002	-.001
	B x C	.001	.006	.002	.000	.001	.004
	A x B x C	.001	.007	.002	.001	.011	.003
	Subject I.D.	.230	--	.471	.224	--	.498
	Error	.119	--	.243	.099	--	.219

Note. CV N = 9423; LV N = 700. Missing data not imputed.

^aNegative values suggest that the model overexplains the total variance.

Table 3.28

Computer-Administered Tests: Analysis of Variance Due to Item Parameters^a--
Target Tracking Test 2

Log Distance : + 1	Variance Components	Concurrent Validation			Longitudinal Validation		
		Variance	Prop. of W/In	Prop. of Total	Variance	Prop. of W/In	Prop. of Total
	Total Variance	.488	--	1	.496	--	1
	Within Subject Variance	.095	1	.195	.100	1	.201
	A. Crosshair Speed	.066	.691	.134	.068	.685	.138
	B. Speed Ratio	.025	.265	.052	.027	.275	.055
	C. Number of Turns	.000	.000	.000	.000	.002	.000
	A x B	-.002	-.016	-.003	-.001	-.015	-.003
	A x C	.000	-.003	-.001	.000	-.002	.000
	B x C	-.002	-.024	-.005	-.002	-.021	-.004
	A x B x C	.008	.086	.017	.007	.075	.015
	Subject I.D.	.255	--	.524	.270	--	.544
	Error	.138	--	.282	.126	--	.255

Note. CV N = 9423; LV N = 700. Missing data not imputed.

^aNegative values suggest that the model overexplains the total variance.

Table 3.29

**Computer-Administered Tests: Analysis of Variance Due to Item Parameters^a--
Target Shoot Test**

		Concurrent Validation			Longitudinal Validation		
Variance Components		Variance ^b	Prop. of W/In	Prop. of Total	Variance ^b	Prop. of W/In	Prop. of Total
Log Distance: + 1							
A.	X-hair Speed	.014	.338	.031	.015	.374	.034
B.	Speed Ratio	.006	.139	.013	.004	.105	.010
C.	Item Difficulty ^c	.015	.359	.033	.015	.377	.034
	A x B	.002	.041	.004	.001	.024	.002
	A x C	.000	.008	.001	.001	.017	.002
	B x C	.000	-.007	-.001	.000	-.009	-.001
	A x B x C	.005	.121	.011	.004	.112	.010
	Subject I.D.	.045	--	.097	.029	--	.066
	Error	.376	--	.812	.372	--	.843
	Total Variance	.464	--	1	.441	--	1
	Within Subject Variance	.043	1	.092	.040	1	.091
Time to Fire:							
A.	X-hair Speed	45.15	.045	.003	-41.71	-.042	-.003
B.	Speed Ratio	106.11	.106	.008	125.67	.127	.010
C.	Item Difficulty	392.76	.392	.029	371.55	.376	.029
	A x B	-42.83	-.043	--	-45.56	-.046	-.004
	A x C	-61.89	-.062	--	54.37	.055	.004
	B x C	-58.99	-.059	--	-66.07	-.067	-.005
	A x B x C	620.65	.620	.047	589.49	.597	.046
	Subject I.D.	1888.17	--	.142	1909.22	--	.150
	Error	10434.62	--	.783	9819.78	--	.772
	Total Variance	13323.75	--	1	12716.76	--	1
	Within Subject Variance	1000.96	1	.075	987.75	1	.078
Proportion of Hits:							
A.	X-hair Speed	.006	.317	.024	.006	.355	.025
B.	Speed Ratio	.003	.136	.010	.002	.113	.008
C.	Item Difficulty	.007	.378	.029	.006	.362	.025
	A x B	.000	.019	.001	.000	-.020	-.001
	A x C	.000	.014	.001	.001	.058	.004
	B x C	.000	--	--	-.001	-.030	-.002
	A x B x C	.003	.142	.011	.003	.162	.011
	Subject I.D.	.011	--	.042	.011	--	.044
	Error	.226	--	.882	.227	--	.886
	Total Variance	.256	--	1	.256	--	1
	Within Subject Variance	.019	1	.076	.018	1	.070

Note. CV N = 9423; LV N = 700. Missing data not imputed.

^aNegative values suggest that the model overexplains the total variance.

^bTime to Fire in tenths of seconds.

^cLevels of item difficulty were based on segment lengths and number of turns.

Table 3.30

**Computer-Administered Tests: Analysis of Variance Due to Item Parameters^a--
Cannon Shoot Test**

		Concurrent Validation			Longitudinal Validation		
Variance Components ^b		Variance ^c	Prop. of W/In	Prop. of Total	Variance ^c	Prop. of W/In	Prop. of Total
Mean Absolute Time Score:							
A. DISTIMCN		220.94	.586	.117	233.36	.662	.131
B. Angle		-8.37	-.022	-.004	-17.14	-.049	-.010
A x B		164.57	.436	.087	136.53	.387	.077
Subject I.D.		50.47	--	.027	47.68	--	.027
Error		1458.47	--	.773	1380.79	--	.775
Total Variance		1886.09	--	1	1781.22	--	1
Within Subject Variance		377.14	1	.200	352.74	1	.198
Mean Absolute Time Score:							
A. DISTOFIM		-21.27	-.117	-.012	-4.14	-.026	-.002
B. Angle		-32.16	-.176	-.018	-37.75	-.241	-.022
A x B		235.95	1.293	.131	198.27	1.268	.116
Subject I.D.		47.22	--	.026	44.33	--	.026
Error		1575.61	--	.873	1501.34	--	.882
Total Variance		1805.35	--	1	1702.05	--	1
Within Subject Variance		182.52	1	.101	156.38	1	.092
Mean Absolute Time Score:							
A. DISTBGOF		144.92	.455	.078	117.99	.426	.068
B. Angle		-21.94	-.069	-.012	-41.32	-.149	-.024
A x B		195.82	.614	.105	200.20	.723	.115
Subject I.D.		49.53	--	.027	46.61	--	.027
Error		1492.40	--	.802	1419.28	--	.814
Total Variance		1860.75	--	1	1742.76	--	1
Within Subject Variance		318.81	1	.171	276.87	1	.159
Mean Absolute Time Score:							
A. DISTBGOF	No				132.28	.312	.072
B. DISTIMCN					236.90	.558	.130
A x B	CV				55.08	.130	.030
Subject I.D.					48.38	--	.026
Error	Data				135.53	--	.741
Total Variance					1828.17	--	1
Within Subject Variance	Available				424.26	1	.232

Note. CV N = 9423; LV N = 700. Missing data not imputed.

^aNegative values suggest that the model overexplains the total variance.

^bDISTIMCN = distance from the cannon to the impact point; DISTOFIM = distance from the target to the impact point at the optimal fire point; DISTBGOF = distance from the point of target onset to the optimal fire point.

^cMean Absolute Time Score in tenths of seconds.

Table 3.31

Computer-Administered Tests: Analysis of Variance Due to Item Parameters -- Target Identification Test

Variance Components	Longitudinal Validation ^a	
	Variance	Proportion of Total Variance
Proportion Correct		
Difficulty	.01	.05
Subject I.D.	.01	.06
Error	.08	.89
Total Variance	.09	1.00
Within-Subject Variance	.01	.05
Decision Time ^b		
Difficulty	848.89	.08
Subject I.D.	3652.53	.34
Error	6121.70	.58
Total Variance	10623.12	1.00
Within-Subject Variance	848.89	.08

Note. N = 700. Missing data not imputed.

^aThe comparison shown here was not made during the Concurrent Validation

^bIn tenths of seconds.

Table 3.32

Computer-Administered Tests: Analysis of Variance Due to Item Parameters -- Number Memory Operation Time^a

Variance Components	Concurrent Validation ^a	
	Variance	Proportion of Total Variance
Operations Type ^b		
Subject I.D.	1199.30	.04
Error	5324.39	.19
Total Variance	21458.11	.77
Within-Subject Variance	27981.80	1.00
	1199.30	.04

Note. N = 9423. Missing data not imputed.

^aIn tenths of seconds.

^bCell means and standard deviations (in hundredths of seconds) for the four types of operations were: Addition, M = 184.69, SD = 128.41; subtraction, M = 224.78, SD = 165.55; multiplication, M = 254.25, SD = 174.68; and division, M = 258.29, SD = 183.90.

for all of the within-subject variance because it is the only within-subject parameter. It accounts for approximately 8 percent of the total variance. This same comparison was not made during Concurrent Validation, so Table 3.31 provides only the results from the subsample of longitudinal Initial sample data.

The Number Memory Test is scored by parameters in accordance with conceptual underpinnings that drove its development (Peterson, et al., 1987). As mentioned before, it was not included in analyses because its scoring (parameter-wise) was not under reassessment and because of the need to conserve resources.

Previous research on tests like Number Memory suggests that the type of operation (addition, subtraction, multiplication, and division) influences operation time (Toquam, Corpe, & Dunnette, 1986). During the field test and during Concurrent Validation, we conducted an analysis of variance to investigate this effect (Table 3.32 shows Concurrent Validation results; field test results are provided in Peterson, 1987). As with the Target Identification Test, the type of operation parameter accounts for all of the within-subject variance because it is the only within-subject parameter. It accounts for approximately 4 percent of the total variance. Moreover, the means and standard deviations of operation time for the different types of operations illustrate the impact of the type of operation on the time scores (see Table 3.32). The mean operation time for addition was 184.69 hundredths of seconds (SD=128.41) compared to 258.29 for division (SD=183.90). For these reasons, the Number Memory Test was scored within parameters during Concurrent Validation. Because the data collected previously support the scoring of operation time by the type of operation, we concluded this scoring routine should be retained for Longitudinal Validation.

Implications for Scoring

The analysis of variance results demonstrate that examinees' scores on particular items are influenced, sometimes to a great degree, by the parameters of those items. The major reason this finding is so important is that we expect each individual to have some missing time scores on the perceptual tests.³ If the examinee has missing scores for only the more difficult items, a mean time score computed across all items will be weighted in favor of the easier items, and vice versa. For example, as shown in Figure 3.3, decision time on Perceptual Speed and Accuracy increases with the number of characters in the item. If a test-taker misses or "times-out" (i.e., does not respond to an item before the time limit is exceeded) on many of the difficult items, his or her reaction time pooled across all items without regard to parameters might appear "fast". It is also important to note here that examinees are less likely to get the "harder" items correct; that is, proportion correct decreases as the number of characters in the stimulus increases. Importantly, this will result in more missing data for

³Missing time score data may occur for any of several reasons: if the subject "timed-out" on the item (i.e., does not respond to an item before the time limit is exceeded), if the subject answered the item incorrectly, if the data were for some reason unreadable on the disk, etc.

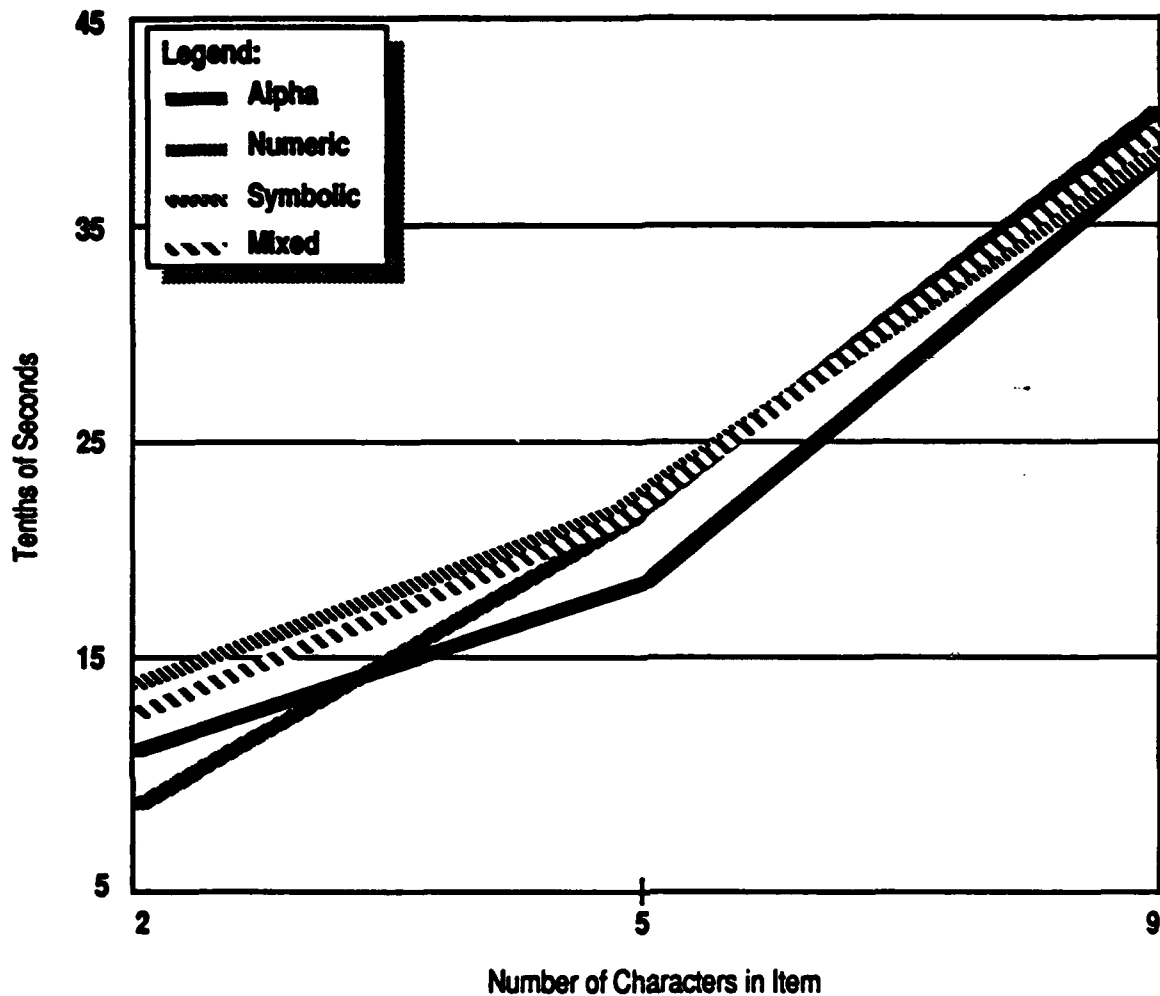


Figure 3.3. Perceptual Speed and Accuracy Test decision time means by parameters.

the harder items because incorrect items are not used to compute decision time.

We concluded that decision time scores should first be computed within levels of the major parameter affecting decision time. Next, means of means should be computed for the test score. Taking the mean within each parameter level and computing the mean of means as a final score ensures equal weight to items at different parameter levels. Also, methods of trimming data (i.e., excluding responses when computing test scores) should be applied within parameter levels to avoid inappropriately flagging responses to the most difficult and easiest items as outliers.

The variance components analyses suggest that movement time (measured on the reaction time tests) is not influenced, or is minimally influenced, by the parameters that impact decision time. We concluded that movement time scores may, therefore, be pooled without regard to parameter levels.

The psychomotor tests have no item-based time limit, no proportion correct score, and no missing data. We are, therefore, not concerned about how missing data will affect the test scores because the mean computed across all items is equivalent to the mean of means taken within parameter, as long as there are equal numbers of items within parameter levels. This holds true for Target Tracking 1 and Target Tracking 2, and it is not necessary to consider parameters in scoring these tests. Target Shoot, however, has different numbers of items for each "difficulty" parameter level, and we have (below) examined alternate ways of scoring this test. Cannon Shoot within-subject variance is influenced by too many parameters and interactions to allow simple classification of items into categories; it should, therefore, be scored without regard to item parameters.

Analysis of Possible Scoring Methods/Changes

The main goal of this phase was to examine the psychometric properties of alternate scoring methods. For example, Roznowski (1987) had found that median reaction time scores had greater test-retest reliabilities than means on simple and choice reaction time tests. There has, however, been little other documented, systematic research on alternate scoring methods, and opinions are mixed. Philip Ackerman (personal communication, 1990) prefers to use mean reaction times and suggests that inclusion of aberrant responses may enhance the validity of the measure. Other researchers have used the median or have removed aberrant responses before computing the mean.

Reaction Time Tests: Scoring Procedures

Three alternate methods of scoring decision time (DT) were examined for each reaction time test (Simple Reaction Time, Choice Reaction Time, Perceptual Speed and Accuracy, Short-Term Memory, Target Identification, and Number Memory). The methods were:

- The median.
- The "clipped" mean (the mean decision time after elimination of the examinee's highest and lowest DT).
- The mean after applying a three-standard deviation data elimination rule: the mean DT computed after deleting DTs that are three standard deviations (the examinee's within-test standard deviation) outside the examinee's untrimmed mean.

For tests having critical item parameters (i.e., Perceptual Speed and Accuracy, Short-Term Memory, Target Identification, and Number Memory), alternate scores were computed within parameter levels and means were taken across parameter levels. For example, the Perceptual Speed and Accuracy median score is the mean taken across the median DT for two-character items, the median DT for five-character items, and the median DT for the nine-

character items. We also examined median, clipped, and 3SD rule alternate scores for movement time on these tests.

Reaction Time Tests: Results and Conclusions

Median decision time scores were "faster" than the other scores and had slightly less variance. For example, for the Simple Reaction Time Test, the median Decision Time Score was 28.61 hundreds of seconds against 29.57 for the clipped mean, and 30.98 for the 3SD mean; the standard deviations were 8.57, 10.05, and 12.35, respectively. For the Short-Term Memory Test, the median Decision Time score was 79.01 hundreds of seconds, against 81.05 for the clipped mean and 84.08 for the 3SD mean; the standard deviations were 21.95, 22.55, and 24.79, respectively. The three-standard-deviation rule did not eliminate any data when there were very few items on the test or within each parameter level. Means from the rule were equal to untrimmed means.

Table 3.33 provides uniqueness estimates, and split-half and test-retest reliability estimates for the alternate scores. In general, the clipping procedure resulted in higher split-half reliabilities than did other procedures. Alternate scoring procedures have the greatest impact on Simple Reaction Time DT, where the median is the most reliable score. For the other tests, the clipping procedure usually resulted in better reliability; there is, however, no large difference in reliability across alternate methods.

With regard to movement time, the split-half reliability for the median is greater than that for clipped and 3SD rule means; the test-retest reliability of the mean of the median MT is highest, .73. All of the alternate scores produce test-retest reliabilities greater than that reported for pooled movement time during Concurrent Validation, .66.

We concluded that the median decision time score should be used for the Simple and Choice Reaction Time tests, mainly because of the increase in test-retest reliability for the Simple Reaction Time Test. Although no large improvements in reliabilities with alternate scores were observed for the other reaction time tests, the clipped mean procedure should be used because it tends to have slightly better qualities overall. For movement time, the pooled median movement time is clearly the best score.

Target Shoot Test: Scoring Procedures

As noted, only one alternate method of scoring needed to be examined for the psychomotor tests. We analyzed pooled Target Shoot scores computed within item difficulty (based on segment lengths and numbers of turns).

Target Shoot Test: Results and Conclusions

Scoring the Target Shoot Test within parameter level makes the test more difficult (i.e., increases mean log [distance + 1] and time to fire). Table 3.34 provides uniqueness estimates and split-half and test-retest

Table 3.33

Computer-Administered Reaction Time Tests: Reliability Coefficients, Squared Multiple Regression Coefficients (vs. All ASVAB Subtests), and Uniqueness Estimates

Test	R ²	Split-Half Reliability ^a	Uniqueness ^b (Split-Half)	Test-Retest Reliability ^c	Uniqueness ^d (Test/Retest)
Simple Reaction Time ^e					
Median Decision Time	.01	NC	NC	.40	.39
Clipped Decision Time	.01	.82	.81	.32	.31
Mean Decision Time After 3SD					
Rule is Applied	.01	.77	.76	.20	.19
Proportion Correct	.01	.68	.67	.00	---
Choice Reaction Time					
Median Decision Time	.05	.94	.89	.68	.63
Clipped Decision Time	.05	.96	.91	.70	.65
Mean Decision Time After 3SD					
Rule is Applied	.05	.96	.91	.70	.65
Proportion Correct	.01	.65	.64	.23	.22
Short-Term Memory					
Median Decision Time	.05	.94	.89	.65	.60
Clipped Decision Time	.05	.97	.92	.66	.61
Mean Decision Time After 3SD					
Rule is Applied	.05	.93	.88	.63	.58
Proportion Correct	.05	.67	.62	.36	.31
Perceptual Speed and Accuracy					
Median Decision Time	.05	.93	.88	.65	.60
Clipped Decision Time	.05	.96	.91	.66	.61
Mean Decision Time After 3SD					
Rule is Applied	.05	.95	.90	.63	.58
Proportion Correct	.03	.67	.64	.49	.46
Target Identification					
Median Decision Time	.13	.95	.82	.75	.62
Clipped Decision Time	.14	.96	.82	.75	.61
Mean Decision Time After 3SD					
Rule is Applied	.14	.95	.81	.75	.61
Proportion Correct	.03	.75	.72	.34	.31
Pooled Movement Time ^f					
Mean of Medians	.05	.98	.93	.73	.68
Means of Clipped Means	.05	.92	.87	.72	.67
Means of Means After 3SD					
Rule is Applied	.04	.94	.90	.69	.65
Number Memory					
Final Response Time					
Median	.21	.85	.64	.60	.39
Clipped Mean	.23	.90	.67	.61	.38
Mean After 3SD Rule is Applied	.23	.89	.66	.60	.37
Input Time					
Median	.10	.91	.81	.47	.37
Clipped Mean	.11	.94	.83	.47	.36
Mean After 3SD Rule is Applied	.11	.93	.82	.46	.35

(Continued)

Table 3.33 (Continued)

Computer-Administered Reaction Time Tests: Reliability Coefficients, Squared Multiple Regression Coefficients (vs. All ASVAB Subtests), and Uniqueness Estimates

Test	R ²	Split-Half Reliability ^a	Uniqueness ^b (Split-Half)	Test-Retest Reliability ^c	Uniqueness ^d (Test/Retest)
Number Memory (Continued)					
Pooled Operations Time					
Mean of Medians	.27	.92	.65	.70	.43
Means of Clipped Means	.27	.95	.68	.72	.45
Means of Means After 3SD Rule is Applied	.28	.93	.65	.72	.44
Proportion Correct	.17	.72	.55	.57	.40

^aLV Initial Sample, N = 6909-6984.

^bUniqueness = Split-Half Reliability - R²

^cCV Test-Retest Sample, N = 472-479.

^dUniqueness = Test-Retest Reliability - R²

^eFollowing on introductory practice items, the Simple Reaction Time Test has only 10 scorable items, too few for computing the split-half median.

^fMovement Time is pooled across five tests: Simple Reaction Time, Choice Reaction Time, Short-Term Memory, Perceptual Speed and Accuracy, and Target Identification.

Table 3.34

Alternate Scores for Target Shoot Test: Reliability and Squared Multiple Regression Coefficients (vs. All ASVAB Subtests), and Uniqueness Estimates

Test	R ²	Split-Half Reliability ^a	Uniqueness ^b (Split-Half)	Test-Retest Reliability ^c	Uniqueness ^d (Test/Retest)
Target Shoot					
Mean Log (Distance + 1)	.06	.81	.75	.41	.35
Mean Time to Fire	.07	.88	.81	.60	.53
Mean of Mean Log (Distance + 1)	.06	.77	.71	.42	.36
Mean of Mean Time to Fire	.07	.84	.77	.58	.51
Hit (Miss + No Fire) Proportion	.07	.66	.59	.48	.41

^aLV Initial Sample, N = 6979.

^bUniqueness = Split-Half Reliability - R²

^cCV Test-Retest Sample, N = 472 - 479.

^dUniqueness = Test-Retest Reliability - R²

reliability estimates for the alternate scores. The split-half reliability for the "mean of mean" Target Shoot scores is slightly lower than for their "whole test" counterparts. Scoring Target Shoot trials within difficulty level increased the test-retest reliability of the log (distance + 1) score slightly and decreased the time score reliability slightly.

The psychometric differences between the alternate scores were too inconsequential to favor either scoring method. We chose to use the mean of mean scores on Target Shoot because it is conceptually desirable to give equal weight to the different levels of item difficulty (such as target and cross hair speed). The mean log (distance + 1) increases from 2.17 to 2.21 when computed by difficulty and the mean time to fire rises from 22.6 to 23.3 tenths of seconds (N = 6929).

Final Screening Rules

After the procedures for computing basic scores were decided, a set of final screening rules was developed to ensure high-quality data without unnecessary deletion of responses.

First, a minimum criterion for the number of non-missing items was set for each test. Missing items could occur for several reasons: The examinee had never taken the test; the items had been identified as missing during the first cleaning pass through the data; the data for an item had not been readable on the disk; the examinee had "timed out," that is, not responded to the item before time ran out; the examinee had not "fired" at a target; or, for the six perceptual tests, the item was answered incorrectly.

An examination of frequency distributions of missing data found that most examinees had complete or almost complete data, even though the cleaning rules produced designations of missing data. For example, 86 percent of the Initial sample examinees had complete data on the Simple Reaction Time test after data were cleaned, and another 10 percent had data for nine of the ten items on this test. Likewise, on Choice Reaction Time, 88 percent of the examinees had data for at least 29 of the 30 items, and another 10 percent had data for at least 26 out of 30 items. Similarly, for the other tests we found that most examinees had very little missing data.

The minimum number of items required was 90 percent for the four psychomotor tests. The percentage of examinees who fell below the criterion for each of these tests is as follows:

	<u>Initial Sample</u>	<u>Sample 2</u>
Target Tracking 1	0.1%	0.2%
Target Tracking 2	0.2%	0.3%
Target Shoot	3.6%	3.9%
Cannon Shoot	0.7%	0.9%

For Simple and Choice Reaction Time tests, we applied a 50 percent criterion -- that is, examinees had to respond correctly to at least 50 percent of the items on these tests -- and fewer than 1 percent of the

examinees fell below the criterion on each test. For Perceptual Speed and Accuracy, Target Identification, Short-Term Memory, and Number Memory, two criteria were applied. First, because means of "clipped" means are computed within cells for each of these tests, at least three responses are required for each cell. Second, we employed a chance-level performance criterion for each test. For tests that have two response options (Perceptual Speed and Accuracy, Short-Term Memory, and Number Memory), this criterion was 50 percent. For Target Identification, which has three response alternatives, the criterion was set at 33 percent. The percentage of examinees who fell below the criterion for each of these tests is shown below:

	<u>Initial Sample</u>	<u>Sample 2</u>
Perceptual Speed and Accuracy	0.5%	0.5%
Short Term Memory	1.1%	1.1%
Target Identification	1.1%	1.5%
Number Memory	1.4%	1.4%

To analyze the overall impact of the screening rules, we applied all the rules and deleted examinees who had a computer test score missing. Out of the Initial sample of 7,000, 6,565 (94 percent) examinees were retained. For Sample 2, 6563 cases were retained out of 7,000 (94 percent).

Basic Scores for Further Analysis

After alternate scores were compared, 20 basic scores were selected for further analysis. The means, standard deviations, and reliabilities of the cognitive/perceptual test scores are provided in Tables 3.35 and 3.36. The tables provide data from Concurrent Validation, the Initial Longitudinal sample, and Longitudinal Sample 2, after screening rules are applied.

As shown, data from the two longitudinal samples are highly similar. The differences between means and standard deviations from the concurrent sample and those from longitudinal samples probably result from the detailed scoring changes made during the current scoring effort. The Decision Mean Time scores all show extremely high split-half (or within-testing-session) reliabilities, and adequate test-retest reliabilities, except for Simple Reaction Time, which has a low test-retest reliability. This was to be expected as the Simple Reaction Time test has few items and largely serves to acquaint the examinee with the testing apparatus.

The Proportion Correct scores show moderate split-half reliabilities and low to moderate test-retest reliabilities. The lower reliabilities for Proportion Correct scores were expected since these tests were designed to produce the most variance for decision time, with relatively low variance for proportion correct. That is, ample time was allowed for examinees to make a response to each item (9 seconds). As Table 3.35 shows, that is what occurred -- a very high proportion of correct scores with relatively small variance (again, compared to decision time scores).

The means, standard deviations, reliabilities, and uniqueness estimates for the psychomotor test scores are given in Tables 3.37 and 3.38. Note that the means and standard deviations of the scores are very similar across the

Table 3.35

Means of Computer-Administered Cognitive/Perceptual Measures From Concurrent Validation and Longitudinal Validation Samples

Measure	Trial Battery (Concurrent Validation) (N = 9099-9274)		Experimental Battery ^b (Longitudinal Validation)			
	Mean ^a	SD	Initial Sample (N = 6436)		Sample 2 (N = 6435)	
			Mean	SD	Mean	SD
Simple Reaction Time: Decision Time Mean	31.84	14.82	28.43	7.68	28.27	7.19
Simple Reaction Time: Proportion Correct	.98	.04	.98	.05	.98	.05
Choice Reaction Time: Decision Time Mean	40.93	9.77	38.55	7.70	38.45	8.35
Choice Reaction Time: Proportion Correct	.98	.03	.98	.04	.98	.03
Perceptual Speed & Accuracy: Decision Time Mean	236.91	63.38	227.11	62.94	227.29	63.95
Perceptual Speed & Accuracy: Proportion Correct	.87	.08	.86	.08	.86	.08
Short Term Memory: Decision Time Mean	87.72	24.03	80.48	21.90	80.91	22.30
Short Term Memory: Proportion Correct	.89	.08	.89	.07	.89	.07
Target Identification: Decision Time Mean	193.65	63.13	179.42	59.60	178.08	59.80
Target Identification: Proportion Correct	.91	.07	.90	.08	.90	.09
Number Memory: Input Response Time Mean	142.84	55.24	141.28	52.63	140.84	52.60
Number Memory: Operations Time Pooled Mean	233.10	79.71	208.73	74.58	207.71	75.43
Number Memory: Final Response Time Mean	160.70	42.63	152.83	41.90	152.56	41.39
Number Memory: Proportion Correct	.90	.09	.86	.10	.86	.10
Pooled Mean Movement Time ^c	33.61	8.03	28.19	6.06	28.37	6.19

^aMean response time values are reported in hundredths of seconds.^bExperimental Battery data were screened for missing data before reliabilities were computed.^cMovement Time is pooled across Simple Reaction Time, Choice Reaction Time, Short-Term Memory, Perceptual Speed and Accuracy, and Target Identification.

Table 3.36

Reliability Estimates for Computer-Administered Cognitive/Perceptual Test Scores

Measure	Split-Half Estimates			Test-Retest Estimates	
	TB (N = 9099-9274)	EB ^a Initial Sample (N = 6215)	EB ^a Sample 2 (N = 6096)	TB (N = 473-479)	TB Rescored with EB Scoring
Simple Reaction Time: Decision Time Mean	.88	.83	.87	.23	.40
Simple Reaction Time: Proportion Correct	.46	.50	.53	.02	.00
Choice Reaction Time: Decision Time Mean	.97	.93	.94	.69	.68
Choice Reaction Time: Proportion Correct	.57	.58	.54	.23	.23
Perceptual Speed & Accuracy: Decision Time Mean	.94	.96	.96	.63	.66
Perceptual Speed & Accuracy: Proportion Correct	.65	.62	.61	.51	.49
Short Term Memory: Decision Time Mean	.96	.97	.97	.66	.66
Short Term Memory: Proportion Correct	.60	.50	.48	.41	.36
Target Identification: Decision Time Mean	.97	.97	.97	.78	.76
Target Identification: Proportion Correct	.62	.66	.69	.40	.34
Number Memory: Input Response Time Mean	.95	.94	.95	.47	.47
Number Memory: Operations Time Pooled Mean	.93	.95	.95	.73	.72
Number Memory: Final Response Time Mean	.88	.90	.90	.62	.61
Number Memory: Proportion Correct	.59	.59	.58	.53	.57
Pooled Mean Movement Time ^b	.74	.97	.97	.66	.73

^aThese reliabilities are for scores computed by the final Longitudinal Validation (Experimental Battery) scoring program.

^bExperimental Battery data were screened for missing data before reliabilities were computed.

^cMovement Time is pooled across Simple Reaction Time, Choice Reaction Time, Short-Term Memory, Perceptual Speed and Accuracy, and Target Identification.

Table 3.37

Means of Computer-Administered Psychomotor Tests From Concurrent Validation and Longitudinal Validation Samples

	Concurrent Validation (N = 8892-9251)		Longitudinal Validation			
	Mean ^a	SD	Initial Sample (N = 6436)		Sample 2 (N = 6435)	
			Mean	SD	Mean	SD
Target Tracking 1						
Mean Log (Distance + 1)	2.98	.49	2.89	.46	2.89	.47
Target Tracking 2						
Mean Log (Distance + 1)	3.70	.51	3.55	.52	3.55	.52
Target Shoot						
Mean Log (Distance + 1)	2.17	.24	2.20	.23	2.19	.23
Mean Time-to-Fire	235.39	47.78	230.98	50.18	231.01	50.67
Cannon Shoot						
Mean Absolute Time Discrepancy	43.94	9.57	44.03	9.31	44.10	9.40

^aTime-to-fire and time-discrepancy measures are in hundredths of seconds. Logs are natural logs.

Table 3.38

Reliability and Uniqueness Estimates for Computer-Administered Psychomotor Test Scores

	Split-Half Estimates			Test-Retest Estimates		Uniqueness	
	TB (N = 9099-9274)	EB Initial Sample (N = 6215)	EB ^a Sample 2 (N = 6096)	EB ^a TB (N = 473-479)	TB Rescored with EB Scoring	Split- Half	Test- Retest
Target Tracking 1							
Mean Log (Distance + 1)	.98	.98	.98	.74	b	.80	.56
Target Tracking 2							
Mean Log (Distance + 1)	.98	.98	.98	.85	b	.76	.63
Target Shoot							
Mean Log (Distance + 1)	.74	.73	.72	.37	.42	.71	.36
Mean Time-to-Fire	.85	.84	.84	.58	.58	.77	.51
Cannon Shoot							
Mean Absolute Time Discrepancy	.65	.64	.65	.52	b	.57	.39

^aThese reliabilities are for scores computed by the final Longitudinal Validation (Experimental Battery) scoring program. Experimental Battery data were screened for missing data before reliabilities were computed.

^bTB and EB scoring methods are the same for this test.

three samples. The split-half reliabilities are uniformly high. The test-retest correlations are high for the two tracking test scores, but are low to moderate for Cannon Shoot and Target Shoot. Even so, there remains a large amount of unique reliable variance for predicting performance.

Comparison of Gender and Race Subgroup Scores

Mean scores by gender and by race were computed to investigate gender and race differences on the tests. We also computed effect sizes, the standardized mean difference between the two subgroup means.

Gender Differences

Table 3.39 shows means, standard deviations, and effect sizes for gender subgroups from the CV, LV Initial, and LV Sample 2 samples. Positive effect sizes (d) indicate greater mean performance by men, and negative values reflect better performance by women.

For the decision time measures for the most part, effect sizes fluctuated between 0 to .20, usually favoring men. For Target Identification decision time, we consistently found about a one-half standard deviation difference in means, favoring men.

On the proportion correct scores, almost all of the effect sizes favored women. The largest effect was on Perceptual Speed and Accuracy proportion correct where women consistently outperformed men by over one-third of a standard deviation.

On the psychomotor test scores, effect sizes all favored men. The Target Shoot time score had the smallest effect (one-half standard deviation), and the largest differences (over one and one-quarter standard deviation) were observed for the two tracking tests. Similarly, men performed better on movement time than women, by about a one-half standard deviation.

Race Differences

Means, standard deviations, and effect sizes for race subgroups are shown in Table 3.40. Effect sizes are presented for each minority compared to whites. Positive effect sizes indicate higher mean performance for whites; a negative value indicates superior performance by the comparison group.

For the decision time measures, effect sizes ranged from zero to about two-thirds of a standard deviation for blacks compared to whites and from zero to one-third of a standard deviation for Hispanics compared to whites. On the psychomotor tests, differences were about one-half standard deviation for blacks compared to whites and about one-third standard deviation for Hispanics compared to whites. For the proportion correct scores, differences were about one-tenth standard deviation for blacks compared to whites.

Table 3.39

Computer-Administered Tests: Means and Effect Sizes by Gender (Concurrent, Longitudinal Initial, and Longitudinal Sample 2 Samples)

Measure	Male			Female			Effect Size ^b (d)
	N	Mean ^a	SD	N	Mean ^a	SD	
Perceptual Tests: Time Scores							
Simple Reaction Time DT							
Concurrent Validation	8359	31.71	14.68	884	32.94	15.72	.08
Longitudinal Initial	6072	28.48	8.78	894	29.24	5.79	.09
Longitudinal Sample 2	6076	28.27	7.22	889	29.56	9.88	.17
Choice Reaction Time DT							
Concurrent Validation	8371	41.04	9.72	886	39.97	10.19	-.11
Longitudinal Initial	6094	38.82	7.96	897	37.99	7.56	-.10
Longitudinal Sample 2	6091	38.73	8.53	895	37.64	7.66	-.12
Short Term Memory DT							
Concurrent Validation	8251	87.73	24.24	886	87.65	22.09	.00
Longitudinal Initial	6038	80.61	22.41	888	83.04	19.82	.11
Longitudinal Sample 2	6033	81.14	22.93	890	82.97	21.63	.08
Perceptual Speed and Acc. DT							
Concurrent Validation	8346	236.37	63.99	886	242.10	57.38	.09
Longitudinal Initial	6069	226.80	64.46	893	238.99	58.32	.19
Longitudinal Sample 2	6072	226.80	65.08	892	238.46	59.10	.18
Target Identification DT							
Concurrent Validation	8215	190.53	61.98	878	223.33	66.23	.52
Longitudinal Initial	6030	178.21	60.58	890	206.95	65.04	.47
Longitudinal Sample 2	6009	176.42	59.80	889	205.08	63.90	.47
Number Memory Final Time							
Concurrent Validation	8210	160.92	43.06	879	158.75	38.36	-.05
Longitudinal Initial	6014	153.35	42.84	890	158.28	41.04	.11
Longitudinal Sample 2	6020	153.32	42.42	886	157.24	40.55	.09
Number Memory Input Time							
Concurrent Validation	8210	143.89	56.01	879	133.03	46.54	-.19
Longitudinal Initial	6014	143.53	55.20	890	136.00	50.85	-.13
Longitudinal Sample 2	6020	142.72	54.35	886	135.35	47.87	-.13
Number Memory Operations Time							
Concurrent Validation	8210	232.64	80.45	879	237.60	72.62	.06
Longitudinal Initial	6013	207.91	76.16	890	224.89	70.94	.22
Longitudinal Sample 2	6018	206.97	76.67	885	221.45	72.75	.19
Perceptual Tests: Proportion Correct							
Simple Reaction Time							
Concurrent Validation	8359	.98	.05	884	.99	.03	-.20
Longitudinal Initial	6072	.98	.05	894	.98	.05	-.02
Longitudinal Sample 2	6076	.98	.05	889	.98	.05	-.04
Choice Reaction Time							
Concurrent Validation	8371	.98	.03	886	.99	.02	-.34
Longitudinal Initial	6094	.98	.03	897	.98	.03	-.18
Longitudinal Sample 2	6091	.98	.03	895	.98	.03	-.13
Short Term Memory							
Concurrent Validation	8251	.89	.08	886	.90	.07	-.12
Longitudinal Initial	6038	.88	.07	889	.89	.06	-.19
Longitudinal Sample 2	6033	.88	.07	890	.90	.06	-.25
Perceptual Speed and Acc.							
Concurrent Validation	8346	.87	.08	886	.90	.07	-.37
Longitudinal Initial	6075	.85	.08	895	.88	.07	-.33
Longitudinal Sample 2	6078	.85	.08	892	.88	.07	-.36
Target Identification							
Concurrent Validation	8215	.91	.08	878	.92	.07	-.12
Longitudinal Initial	6030	.89	.09	890	.89	.08	-.01
Longitudinal Sample 2	6009	.89	.09	889	.89	.09	.00

(Continued)

Table 3.39 (Continued)

Computer-Administered Tests: Means and Effect Sizes by Gender (Concurrent, Longitudinal Initial, and Longitudinal Sample 2 Samples)

Measure	Male			Female			Effect Size ^a (d)
	N	Mean ^a	SD	N	Mean ^a	SD	
Number Memory							
Concurrent Validation	8210	.90	.09	879	.90	.08	.00
Longitudinal Initial	6017	.86	.10	891	.85	.09	.06
Longitudinal Sample 2	6023	.86	.10	886	.85	.09	.02
Psychomotor Test Scores							
Target Tracking 1 Distance Score							
Longitudinal Initial	6096	2.85	.44	897	3.41	.44	1.26
Longitudinal Sample 2	6090	2.84	.45	894	3.43	.45	1.28
Target Tracking 2 Distance Score							
Concurrent Validation	8348	3.65	.51	879	4.11	.38	.92
Longitudinal Initial	6089	3.50	.50	897	4.12	.36	1.26
Longitudinal Sample 2	6086	3.50	.50	894	4.12	.35	1.28
Target Shoot Distance Score							
Concurrent Validation	8106	2.16	.23	774	2.31	.29	.63
Longitudinal Initial	5932	2.17	.21	818	2.37	.28	.88
Longitudinal Sample 2	5915	2.17	.20	809	2.37	.27	.90
Target Shoot Time Score							
Concurrent Validation	8106	233.66	47.65	774	253.86	45.32	.42
Longitudinal Initial	5932	228.16	49.80	818	252.43	48.50	.48
Longitudinal Sample 2	5915	228.05	49.97	809	255.04	50.13	.54
Cannon Shoot Time Score							
Concurrent Validation	8337	43.19	9.05	885	51.03	11.29	.84
Longitudinal Initial	6062	43.27	8.70	887	52.46	12.35	.99
Longitudinal Sample 2	6044	43.33	8.82	891	52.40	12.76	.96
Pooled Movement Time							
Concurrent Validation	8375	33.37	8.03	887	35.84	7.69	.30
Longitudinal Initial	5936	27.94	6.12	879	31.08	5.89	.51
Longitudinal Sample 2	5927	28.05	6.13	878	31.56	6.25	.57

^aTime scores are in hundredths of seconds; distance scores are log (distance + 1).

^bd is the standardized mean difference between male and female means. A negative sign indicates superior performance by females.

Table 3.40

Computer-Administered Tests: Means and Effect Sizes by Race (Concurrent, Longitudinal Initial, and Longitudinal Sample 2 Samples)

	White			Black			Hispanic			Other		
	N	Mean ^a	SD	N	Mean ^a	SD	N	Mean ^a	SD	N	Mean ^a	SD
Measure												
Perceptual Tests: Time Scores												
Simple Reaction Time DT												
Concurrent Validation	6006	31.06	11.55	2548	33.27	18.36	335	34.81	26.18	354	31.61	18.39
Longitudinal Initial	4774	28.70	7.53	1704	28.48	11.16	239	27.49	5.53	238	27.98	5.54
Longitudinal Sample 2	4773	28.42	6.43	1709	28.49	9.64	234	27.90	5.43	233	28.70	13.28
Choice Reaction Time DT												
Concurrent Validation	6011	40.42	8.33	2555	41.68	11.06	336	43.82	15.84	355	41.57	13.53
Longitudinal Initial	4790	38.70	7.53	1713	38.64	8.72	240	38.77	8.16	237	39.52	9.12
Longitudinal Sample 2	4787	38.56	8.19	1715	38.76	9.18	234	37.84	7.28	234	38.62	8.67
Short-Term Memory DT												
Concurrent Validation	5976	86.94	23.23	2483	88.93	25.16	330	93.25	28.03	348	87.39	24.51
Longitudinal Initial	4761	79.95	21.13	1686	83.76	24.65	236	79.78	22.03	232	81.74	20.75
Longitudinal Sample 2	4763	80.53	21.85	1684	84.01	25.40	230	80.85	21.37	230	79.94	21.78
Perceptual Speed and Acc. DT												
Concurrent Validation	5998	236.13	61.47	2544	237.91	67.04	336	248.68	65.49	354	231.84	65.54
Longitudinal Initial	4766	227.94	63.09	1709	230.95	66.48	240	225.10	60.17	236	222.61	62.08
Longitudinal Sample 2	4777	227.96	63.09	1704	230.36	69.14	234	220.11	59.58	233	228.03	62.71
Target Identification DT												
Concurrent Validation	5939	180.19	55.05	2479	222.45	68.64	324	213.08	68.51	351	201.38	71.61
Longitudinal Initial	4757	171.48	55.25	1682	210.87	69.25	236	186.86	61.39	234	182.38	68.05
Longitudinal Sample 2	4740	169.65	54.63	1684	208.35	69.26	228	183.10	58.40	230	185.26	59.98
Number Memory Final Time												
Concurrent Validation	5930	152.63	38.80	2485	179.01	44.93	324	168.37	47.26	350	160.62	41.96
Longitudinal Initial	4748	146.63	38.64	1678	174.14	47.38	237	160.90	39.54	230	152.48	40.18
Longitudinal Sample 2	4746	147.30	39.26	1678	172.01	45.39	233	155.98	39.64	233	153.19	40.39
Number Memory Input Time												
Concurrent Validation	5930	135.34	47.26	2485	158.26	65.68	324	150.03	49.31	350	153.82	76.03
Longitudinal Initial	4748	136.36	48.44	1678	158.53	65.18	237	152.05	69.45	230	144.33	51.90
Longitudinal Sample 2	4746	135.81	46.26	1678	156.93	66.49	233	148.21	59.32	233	147.94	60.96
Number Memory Operations Time												
Concurrent Validation	5930	222.96	75.21	2485	252.78	83.34	324	254.95	80.48	350	245.36	91.47
Longitudinal Initial	4748	201.12	70.88	1677	232.53	82.41	237	229.82	83.72	230	213.55	75.71
Longitudinal Sample 2	4744	199.96	71.52	1677	230.54	84.10	233	226.30	72.89	233	215.06	80.41

(Continued)

Table 3.40 (Continued)

Computer-Administered Tests: Means and Effect Sizes by Race (Concurrent, Longitudinal Initial, and Longitudinal Sample 2 Samples)

Measure	White			Black			Hispanic			Other		
	N	Mean ^a	SD	N	Mean ^a	SD	N	Mean ^a	SD	N	Mean ^a	SD
<u>Perceptual Tests: Proportion Correct Scores</u>												
Simple Reaction Time	6006	.99	.04	2548	.98	.05	335	.98	.05	354	.98	.05
Concurrent Validation	4774	.98	.04	1704	.97	.06	239	.98	.05	238	.97	.06
Longitudinal Initial	4773	.98	.04	1709	.97	.06	234	.97	.05	233	.97	.06
Choice Reaction Time												
Concurrent Validation	6011	.99	.03	2555	.98	.04	336	.98	.03	355	.98	.04
Longitudinal Initial	4790	.98	.03	1713	.97	.04	240	.98	.03	237	.97	.04
Longitudinal Sample 2	4787	.98	.03	1715	.97	.04	234	.97	.04	234	.97	.04
Short Term Memory												
Concurrent Validation	5976	.89	.07	2483	.87	.09	330	.89	.08	348	.89	.08
Longitudinal Initial	4761	.88	.07	1687	.87	.08	236	.89	.06	232	.89	.07
Longitudinal Sample 2	4763	.89	.07	1684	.87	.08	230	.89	.06	230	.89	.05
Perceptual Speed and Accuracy												
Concurrent Validation	5998	.88	.08	2544	.86	.09	336	.88	.07	354	.87	.08
Longitudinal Initial	4773	.85	.08	1710	.84	.08	240	.85	.09	236	.85	.07
Longitudinal Sample 2	4782	.85	.08	1705	.84	.08	234	.85	.08	233	.86	.08
Target Identification												
Concurrent Validation	5939	.91	.07	2479	.90	.08	324	.92	.07	351	.90	.08
Longitudinal Initial	4757	.90	.08	1682	.88	.10	236	.90	.09	234	.89	.10
Longitudinal Sample 2	4740	.89	.08	1684	.88	.10	228	.90	.07	230	.90	.08
Number Memory Test												
Concurrent Validation	5930	.91	.08	2485	.88	.09	324	.89	.09	350	.90	.09
Longitudinal Initial	4751	.87	.09	1679	.83	.10	237	.84	.09	230	.86	.09
Longitudinal Sample 2	4748	.86	.10	1679	.83	.10	233	.85	.09	233	.84	.11

(Cont Inued)

Table 3.40 (Continued)

Computer-Administered Tests: Means and Effect Sizes by Race (Concurrent, Longitudinal Initial, and Longitudinal Sample 2 Samples)

	White			Black			Hispanic			Other			
	N	Mean ^a	SD	N	Mean ^a	SD	N	Mean ^a	SD	N	Mean ^a	SD	Effect Size ^b (d)
Psychomotor Test Scores													
Target Tracking 1 Distance Score													
Concurrent Validation	5999	2.87	.44	2550	3.23	.50	336	3.02	.48	354	3.04	.52	.38
Longitudinal Initial	4788	2.83	.44	1716	3.16	.51	240	2.94	.47	238	2.94	.49	.23
Longitudinal Sample 2	4788	2.84	.45	1713	3.15	.52	234	2.91	.46	233	2.95	.51	.24
Target Tracking 2 Distance Score													
Concurrent Validation	5991	3.56	.49	2548	3.99	.43	336	3.76	.46	352	3.77	.52	.42
Longitudinal Initial	4784	3.47	.49	1713	3.90	.48	240	3.61	.50	238	3.61	.52	.27
Longitudinal Sample 2	4786	3.47	.50	1711	3.88	.48	234	3.59	.49	233	3.58	.51	.21
Target Shoot Distance Score													
Concurrent Validation	5870	2.16	.23	2346	2.22	.26	324	2.15	.20	340	2.18	.26	.08
Longitudinal Initial	4680	2.18	.21	1600	2.24	.28	230	2.18	.27	229	2.19	.19	.04
Longitudinal Sample 2	4678	2.18	.20	1584	2.23	.26	225	2.20	.24	221	2.21	.29	.15
Target Shoot Time Score													
Concurrent Validation	5870	226.80	44.86	2346	254.91	48.54	324	245.35	46.72	340	240.34	50.55	.30
Longitudinal Initial	4680	224.33	47.45	1600	248.84	53.02	230	239.13	52.59	229	236.78	51.50	.26
Longitudinal Sample 2	4678	224.71	48.36	1584	249.74	53.12	225	235.07	50.74	221	234.11	49.93	.19
Cannon Shoot Time Score													
Concurrent Validation	5992	42.41	8.68	2542	47.61	10.86	334	43.51	8.83	354	43.88	9.14	.16
Longitudinal Initial	4759	43.26	9.12	1704	47.93	10.79	239	44.04	9.21	236	43.56	8.39	.03
Longitudinal Sample 2	4758	43.39	9.22	1698	47.82	11.13	232	44.05	9.28	231	43.43	8.78	.00
Pooled Movement Time													
Concurrent Validation	6014	33.02	7.57	2557	35.32	8.85	336	33.05	7.83	355	31.88	7.64	-.15
Longitudinal Initial	4706	27.87	5.87	1639	29.82	6.84	232	28.35	6.26	227	27.62	5.86	-.04
Longitudinal Sample 2	4695	28.19	6.11	1644	29.81	6.55	224	27.44	6.14	226	26.56	5.83	-.26

^aTime scores are in hundredths of seconds; distance scores are log (distance + 1)^bd is the standardized mean difference between two subgroups' scores. All effect sizes are relevant to the white group. A negative value indicates superior performance by the comparison group (black, Hispanic, or other).

Composite Formation

Over the course of Project A and the Career Force Project, we have conducted numerous factor analyses of the computer-administered test scores (e.g., principal components, common factors, with and without spatial test scores and ASVAB subtest scores). In conjunction with the factor analyses, parallel analyses (Humphreys & Montanelli, 1975; Montanelli & Humphreys, 1976) have been used to inform the decision about the number of factors to extract.

Summary of Factor-Analytic Results

To compare the factor structure of the computer-administered test scores across Concurrent Validation and Longitudinal Validation samples, the analyses shown in Tables 3.41-3.44 were performed. Tables 3.41 and 3.42 including only computer test scores show three-factor solutions for Concurrent and Initial Longitudinal validation samples, respectively.

Table 3.41

Concurrent Validation: Factor Analysis^a of Computer-Administered Measures

Measure	Factor 1 Psycho- motor	Factor 2 Perceptual Speed	Factor 3 Perceptual Accuracy	h^2 ^b
Target Tracking 1: Mean Log (Distance + 1)	.83	.08	-.16	.72
Target Tracking 2: Mean Log (Distance + 1)	.81	.11	-.13	.68
Target Shoot: Mean Log (Distance + 1)	.50	-.01	-.20	.30
Cannon Shoot: Mean Absolute Time Discrepancy	.47	.17	-.14	.27
Target Shoot: Mean Time to Fire	.38	.25	.09	.22
Pooled Mean Movement Time	.27	.16	.01	.09
Number Memory: Operations Time Pooled Mean	.01	.70	-.08	.50
Number Memory: Final Response Time Mean	.16	.68	-.10	.50
Number Memory: Input Response Time Mean	.07	.56	-.21	.36
Short Term Memory: Decision Time Mean	.16	.45	.24	.29
Target Identification: Decision Time Mean	.43	.43	.31	.46
Choice Reaction Time: Decision Time Mean	.19	.38	.08	.19
Simple Reaction Time: Decision Time Mean	.15	.19	.00	.06
Perceptual Speed and Accuracy: Proportion Correct	-.01	.12	.67	.46
Perceptual Speed and Accuracy: Decision Time Mean	.16	.46	.58	.58
Target Identification: Proportion Correct	-.05	.16	.50	.28
Number Memory: Proportion Correct	-.05	-.24	.43	.25
Short Term Memory: Proportion Correct	-.13	-.12	.36	.16
Choice Reaction Time: Proportion Correct	-.05	-.04	.23	.06
Simple Reaction Time: Proportion Correct	-.07	-.08	.11	.03
Eigenvalue	2.38	2.32	1.76	6.46

Note: N = 8521.

^aPrincipal factor analysis, initial communality estimate = squared multiple correlation, varimax rotation.

^b h^2 = communality (sum of squared factor loadings) for variables.

Table 3.42

**Longitudinal Validation Initial Sample: Factor Analysis^a of
Computer-Administered Measures**

Measure	Factor 1 Perceptual Speed	Factor 2 Psychomotor	Factor 3 Perceptual Accuracy	h^2 ^b
Perceptual Speed and Accuracy: Decision Time Mean	.64	.16	.37	.56
Number Memory: Final Response Time Mean	.62	.17	-.34	.53
Number Memory: Operations Time Pooled Mean	.62	.02	-.25	.45
Target Identification: Decision Time Mean	.56	.36	.18	.48
Short-Term Memory: Decision Time Mean	.55	.17	.00	.33
Number Memory: Input Response Time Mean	.46	.08	-.36	.35
Choice Reaction Time: Decision Time Mean	.43	.10	-.01	.20
Simple Reaction Time: Decision Time Mean	.28	.15	-.02	.10
Target Tracking 1: Mean Log (Distance + 1)	.15	.85	-.13	.76
Target Tracking 2: Mean Log (Distance + 1)	.17	.82	-.12	.72
Target Shoot: Mean Log (Distance + 1)	.05	.59	-.14	.37
Cannon Shoot: Mean Absolute Time Discrepancy	.15	.53	-.18	.34
Pooled Mean Movement Time (5 tests)	.23	.35	.00	.18
Target Shoot: Mean Time-to-Fire	.29	.34	.04	.20
Perceptual Speed and Accuracy: Proportion Correct	.30	-.02	.61	.46
Number Memory: Proportion Correct	-.14	-.11	.52	.30
Target Identification: Proportion Correct	.23	-.13	.41	.24
Short-Term Memory: Proportion Correct	-.07	-.12	.40	.18
Choice Reaction Time: Proportion Correct	-.02	.01	.22	.05
Simple Reaction Time: Proportion Correct	-.04	-.02	.10	.01
Eigenvalue	2.66	2.56	1.58	6.80

Note: N = 6763.

^aPrincipal factor analysis, initial communality estimate = squared multiple correlations, varimax rotation.

^b h^2 = communality (sum of squared factor loadings) for variable.

In general, the findings in Tables 3.41 and 3.42 are very similar, but with a bit more common variance in the LV sample (6.80 versus 6.46). The Psychomotor and Perceptual Speed factors are larger than the Perceptual Accuracy factor in both samples, but the difference is greater in the LV sample. The variable loadings show essentially the same pattern, except that the Perceptual Speed and Accuracy Decision Time Mean, which has a split loading on the two perceptual factors in both samples, loads highest on Perceptual Speed in the LV sample (which makes the most sense theoretically), rather than highest on Perceptual Accuracy, as it did in the CV sample.

Factor analyses of the computer-administered test scores and ASVAB test scores are provided in Tables 3.43 and 3.44 for Concurrent and Longitudinal Initial samples. Comments similar to those made for Tables 3.41 and 3.42 also apply here. The LV sample shows slightly higher common variance (12.43 versus 11.98), and the three largest factors, General Cognition, Psychomotor, and Number, appear in both solutions and have about the same proportion of the common variance (75%). The variable communalities are very similar; the largest difference is .10 for ASVAB Electronics Information, but the average difference is just .02.

The major difference between the two solutions is the separation of speed (decision time) scores and accuracy (proportion correct) scores into separate factors in the LV sample, whereas in the CV sample these measures combined to form a "simple speed and accuracy" factor and a "complex speed and accuracy" factor. However, some fairly severe split loadings occur for variables across these factors in both solutions (see, for example, Perceptual Speed and Accuracy Decision Time Mean and Target Identification Decision Time Mean). It appears that these tests may not have a highly stable structure when factored with the ASVAB, unlike the case when only the computer-administered test scores are factored.

Several findings have emerged consistently across these and other factor analyses of Pilot Trial Battery, CV, and LV sample data. First, Target Tracking 1 Distance, Target Tracking 2 Distance, Target Shoot Distance, and Cannon Shoot Time Score consistently form one factor--Psychomotor. The Target Shoot Time Score loads on the Psychomotor factor when six or fewer computer test score factors are extracted. When larger numbers of factors are extracted, the Target Shoot Time Score forms its own factor. The communality of the Target Shoot Time Score variable is relatively small, and its reliability estimates are relatively low.

Second, the pooled movement time variable usually has loadings split across three or four factors, although its largest loading is on the Psychomotor factor. (We have usually included a pooled movement time score in factor analyses of basic scores even though it is a composite of movement time scores from five tests.)

Third, in factor analyses that include ASVAB subtests, one cross-method factor emerges, combining computer test scores on Number Memory with ASVAB Math Knowledge and Arithmetic Reasoning subtest scores (Tables 3.43 and 3.44). In factor analyses without the ASVAB subtests, Number Memory scores form their own factor after four or five factors are extracted.

Table 3.43

Concurrent Validation: Factor Analysis^a of Computer-Administered Measures and ASVAB Subtests

Measure	Factor 1 General Cognition	Factor 2 Number	Factor 3 Psychomotor	Factor 4 Complex Speed & Accuracy	Factor 5 Simple Speed & Accuracy	h^2 ^b
ASVAB: General Science	.79	.11	.10	.04	.09	.65
ASVAB: Word Knowledge	.73	.11	.01	.10	.11	.57
ASVAB: Mechanical Comprehension	.69	.07	.32	.03	.00	.59
ASVAB: Electronics Information	.66	.00	.18	.02	.02	.48
ASVAB: Auto/Shop	.65	.06	.23	.00	.03	.48
ASVAB: Paragraph Comprehension	.61	.18	.00	.08	.11	.43
ASVAB: Arithmetic Reasoning	.57	.49	.11	.13	.02	.60
ASVAB: Number Operations	.14	.62	.00	.06	.07	.41
ASVAB: Mathematics Knowledge	.51	.52	.08	.13	.00	.56
ASVAB: Coding Speed	.00	.49	.03	.08	.14	.27
Number Memory: Proportion Correct	.16	.41	.07	.32	.02	.30
Number Memory: Input Response Time Mean	.12	.48	.13	.05	.04	.28
Number Memory: Final Response Time Mean	.17	.58	.20	.16	.19	.48
Number Memory: Operations Time Pooled Mean	.20	.66	.06	.21	.01	.54
Target Tracking 1: Mean Log (Distance + 1)	.18	.02	.81	.07	.08	.71
Target Tracking 2: Mean Log (Distance + 1)	.23	.04	.78	.04	.06	.68
Target Shoot: Mean Log (Distance + 1)	.05	.00	.50	.17	.06	.29
Cannon Shoot: Mean Absolute Time Discrepancy	.13	.11	.47	.03	.09	.27
Target Shoot: Mean Time-to-Fire	.15	.11	.37	.21	.07	.23
Pooled Mean Movement Time	.01	.11	.29	.08	.07	.11
Perceptual Speed & Accuracy: Decision Time Mean	.00	.19	.11	.69	.22	.57
Perceptual Speed & Accuracy: Proportion Correct	.05	.11	.05	.67	.04	.46
Target Identification: Proportion Correct	.10	.01	.06	.51	.00	.28
Target Identification: Decision Time Mean	.26	.15	.36	.47	.23	.50
Short-Term Memory: Proportion Correct	.12	.16	.12	.30	.17	.18
Choice Reaction Time: Proportion Correct	.13	.02	.02	.19	.13	.07
Choice Reaction Time: Decision Time Mean	.02	.18	.13	.14	.55	.37
Simple Reaction Time: Decision Time Mean	.00	.03	.10	.00	.45	.22
Short-Term Memory: Decision Time Mean	.04	.22	.14	.35	.37	.34
Simple Reaction Time: Proportion Correct	.12	.01	.01	.09	.23	.08
Eigenvalue	3.86	2.64	2.48	2.00	1.00	11.98

Note: N = 7119.

^aPrincipal factor analysis, initial communality estimate = squared multiple correlation, varimax rotation.^b h^2 = communality (sum of squared factor loadings) for variables.

Table 3.44

Longitudinal Validation Initial Sample: Factor Analysis^a of Computer-Administered Tests and ASVAB Subtests

Measure	Factor 1 General Cognition	Factor 2 Psychomotor	Factor 3 Number	Factor 4 Perceptual Accuracy	Factor 5 Perceptual Speed	$\sum R^2$ ^b
ASVAB: General Science	.81	-.10	-.04	.07	-.07	.68
ASVAB: Word Knowledge	.75	-.02	-.03	.11	-.08	.58
ASVAB: Electronics Information	.73	-.20	-.02	.04	-.04	.58
ASVAB: Mechanical Comprehension	.69	-.36	-.11	.08	.03	.63
ASVAB: Auto/Shop	.62	-.31	-.09	.01	.12	.50
ASVAB: Arithmetic Reasoning	.56	-.13	-.46	.15	.06	.57
ASVAB: Paragraph Comprehension	.57	.00	-.18	.16	-.10	.39
ASVAB: Mathematics Knowledge	.51	-.06	-.50	.15	-.04	.54
ASVAB: Coding Speed	-.09	-.01	-.44	.15	-.18	.25
ASVAB: Number Operations	-.18	.00	-.60	.10	-.12	.42
Number Memory: Mean of Clip Mean Operations Time	-.27	.10	.65	.16	.11	.54
Number Memory: Clip Mean Final Response Time	-.22	.24	.56	.01	.28	.50
Number Memory: Clip Mean Input Time	-.14	.17	.49	-.02	.10	.30
Perceptual Speed and Accuracy: Proportion Correct	.00	.00	-.07	.66	.17	.47
Target Identification: Proportion Correct	.10	-.10	.03	.47	.11	.25
Number Memory: Proportion Correct	.14	-.13	-.38	.39	-.01	.34
Short-Term Memory: Proportion Correct	.07	-.11	-.14	.38	-.15	.20
Choice Reaction Time: Proportion Correct	.08	.02	-.06	.20	-.05	.05
Choice Reaction Time: Median Decision Time	.07	.00	.01	.12	-.11	.03
Short-Term Memory: Mean of Clip Mean Decision Time	.04	.09	.17	-.01	.55	.34
Perceptual Speed & Accuracy: Mean of Clip Mean DT	-.03	.19	.23	.10	.54	.40
Target Identification: Mean of Clip Mean Decision Time	-.05	.17	.22	.50	.50	.58
Simple Reaction Time: Median Decision Time	.04	.35	.18	.34	.41	.52
Target Tracking 1: Mean Log (Distance + 1)	-.17	.14	.06	-.05	.11	.20
Target Tracking 2: Mean Log (Distance + 1)	-.23	.83	.05	-.08	.11	.74
Target Shoot: Mean of Mean Log (Distance + 1)	-.06	.81	.07	-.05	.08	.72
Cannon Shoot: Mean Absolute Time Discrepancy	-.10	.57	.00	-.17	.12	.37
Mean of Median Movement Times (5 tests)	-.04	.54	.11	-.12	.09	.34
Target Shoot: Mean of Mean Time-to-Fire	-.15	.38	.12	.07	.15	.19
		.36	.16	.18	.10	.22
Eigenvalue	3.93	2.90	2.45	1.63	1.51	12.43

N = 6763.

^aPrincipal factor analysis, initial communality estimate = squared multiple correlations, varimax rotation.^b $\sum R^2$ = communality (sum of squared factor loadings) for variables.

Fourth, the Simple Reaction Time Time Score and Choice Reaction Time Time Score form a factor--Basic Speed, and the Simple Reaction Time Proportion Correct and Choice Reaction Time Proportion Correct form a factor--Basic Accuracy.

Fifth, the Perceptual Speed and Accuracy (PSA) Time Score, PSA Proportion Correct, Target Identification (TID) Time Score, and TID Proportion Correct often load together. But, this result is not necessarily consistent across solutions. For example, the proportion correct scores for these tests form a factor separate from the time scores in the five-factor solution for the LV Initial sample (Table 3.44).

Finally, the Short-Term Memory Time Score loads on the Basic Speed factor, and Short-Term Memory Proportion Correct loads with the more complex perceptual test scores. When larger numbers of factors are extracted, Short Term Memory scores form a factor; score loadings are usually split across two or three factors.

Summary of Computer-Administered Test Score Composites

Based on analysis of the computer-administered test scores, eight computer test composites are recommended (Figure 3.4). Four of these composites can be extracted readily from the factor-analytic findings; they had virtually identical counterparts in CV.

Psychomotor -- Sum these variables: Target Tracking 1 Distance, Target Tracking 2 Distance, Target Shoot Distance, and Cannon Shoot Time Score. The Target Shoot Time Score was dropped because its reliabilities are relatively low and excluding this score will enhance the homogeneity and meaningfulness of the Psychomotor composite; all remaining constituent variables have high loadings on factor.

Number Speed and Accuracy -- Subtract Number Memory Time Score from Number Memory Proportion Correct, or reflect⁴ the time score and add them. For the CV two more Number Memory time scores, Input Time and Final Response Time, were included in this composite. However, including all four scores appears to be unnecessary to produce a reliable composite.

Basic Speed -- Sum Simple Reaction Time Time Score and Choice Reaction Time Time Score.

Basic Accuracy -- Sum Simple Reaction Time Proportion Correct and Choice Reaction Time Proportion Correct.

⁴In combining various types of scores, reflecting (reversing the sign of one score) is used to make all scores run in the same direction. Here, for example, Proportion Correct scores are scaled so that higher scores are "better", while Time scores are scaled so that lower scores are better.

Composite	CV Scores	LV Scores
<u>Paper-and-Pencil Test Scores</u>		
Spatial	Assembling Objects Object Rotation Maze Test Orientation Test Map Test Reasoning Test	Assembling Objects Object Rotation Maze Test Orientation Test Map Test Reasoning Test
<u>Computer-Administered Test Scores</u>		
Psychomotor	Target Tracking 1 Distance Target Tracking 2 Distance Cannon Shoot Time Score Target Shoot Distance Target Shoot Time Score	Target Tracking 1 Distance Target Tracking 2 Distance Cannon Shoot Time Score Target Shoot Distance
Number Speed and Accuracy	Number Memory (Operation DT) Number Memory (PC) Number Memory (Input DT) Number Memory (Final DT)	Number Memory (Operation DT) Number Memory (PC)
Basic Speed	Simple Reaction Time (DT) Choice Reaction Time (DT)	Simple Reaction Time (DT) Choice Reaction Time (DT)
Basic Accuracy	Simple Reaction Time (PC) Choice Reaction Time (PC)	Simple Reaction Time (PC) Choice Reaction Time (PC)
Movement Time	-----	Pooled Movement Time *
Short-Term Memory	-----	Short-Term Memory (PC) Short-Term Memory (DT)
Perceptual Speed	Perceptual Speed & Accuracy (DT) Target Identification (DT) Short-Term Memory (DT)	Perceptual Speed & Accuracy (DT) Target Identification (DT)
Perceptual Accuracy	Perceptual Speed & Accuracy (PC) Target Identification (PC)	Perceptual Speed & Accuracy (PC) Target Identification (PC) Short-Term Memory (PC)

Note: DT = Decision Time, PC = Proportion Correct, * Movement time was not included in CV composites.

Figure 3.4. Comparison of Concurrent Validation and Longitudinal Validation composites.

A review of all the information about the scores showed that two more composites have good support:

Movement Time -- Sum Median Movement Time scores on Simple Reaction Time, Choice Reaction Time, Perceptual Speed and Accuracy, Short-Term Memory, and Target Identification.

Short-Term Memory -- Subtract Short-Term Memory Decision Time from Short-Term Memory Proportion Correct, or reflect the time score and add them.

The pooled movement time variable was not used during CV. However, the LV analyses showed that its internal consistency and test-retest reliability improved substantially when medians were used. Also, it should not be placed in the psychomotor composite because the psychomotor scores involve movement

judgment and spatial ability as well as coordination. Movement time is a more basic measure of movement speed.

For the CV, Short-Term Memory scores were placed in composites with Perceptual Speed and Accuracy scores and Target Identification scores. However, the Short-Term Memory scores simply do not fit well conceptually or theoretically with the other test scores, and they do form a separate empirical factor, when enough factors are extracted.

Finally, two composites for the Perceptual Speed and Accuracy and Target Identification scores are recommended:

Perceptual Speed -- Sum Perceptual Speed and Accuracy Time Score and Target Identification Time Score.

Perceptual Accuracy -- Sum Perceptual Speed and Accuracy Proportion Correct and Target Identification Proportion Correct.

The reasons are that the Perceptual Speed and Accuracy (PSA) Time Score, PSA Proportion Correct, Target Identification (TID) Time Score, and TID Proportion Correct often load together. More specifically, the time scores and proportion correct scores both load positively on the same factor; the correlations between the speed and proportion correct scores for these tests are positive. It should be remembered that these are raw scores that have not been reflected. This suggests that individuals who respond quickly are less accurate and those who respond slowly are more accurate. In contrast, on Short-Term Memory and Number Memory, the speed and proportion correct scores are negatively correlated. The fact that Short-Term Memory speed and proportion correct scores correlate negatively, and PSA and TID speed and proportion correct scores correlate positively is one additional reason for forming a separate Short-Term Memory composite. Given these findings, separate speed and accuracy scores, combined across the two tests, were completed.

Lloyd Humphreys (personal communication, March 1990) has suggested that the correlations between different scores on the same test (e.g., Perceptual Speed and Accuracy Decision Time and Proportion Correct) might be inflated because the multiple scores, initially recorded for a single item, are not independent. A way of removing this dependence is to compute each score on alternate split halves of the test and to use these scores in subsequent factor analyses. Consequently, we prepared a data base in which, for all tests that have two scores, one score was computed using one half of the items and the other score used the alternate set of items.

Correlations between total proportion correct and time scores and between alternate half proportion correct and time scores appear in Table 3.45. The alternate score correlations for Perceptual Speed and Accuracy and Target Identification are lower than those for the total scores, suggesting that item interdependence does inflate the total score correlation somewhat for these two tests. The correlations remained essentially the same for the other four tests. The profile of correlations is strikingly similar for total scores and alternate half scores; speed/proportion correct correlations are positive for Perceptual Speed and Accuracy and Target Identification, negative for Short-Term Memory and Number Memory, and essentially zero for Simple and Choice Reaction Time.

Table 3.45**Longitudinal Validation: Correlations Between Proportion Correct and Time Scores for Total and Alternate Half Scores on Perceptual Computer-Administered Tests**

Test	Correlation Between Proportion Correct and Time Score			
	Total Scores		Alternate Half Scores	
	Initial Sample	Sample 2	Initial Sample	Sample 2
Simple Reaction Time	-.06	-.10	-.05	-.06
Choice Reaction Time	.01	-.02	.00	.00
Perceptual Speed and Accuracy	.50	.50	.40	.40
Target Identification	.22	.24	.13	.16
Short-Term Memory	-.12	-.09	-.12	-.09
Number Memory	-.19	-.19	-.18	-.18

We also compared factor solutions based on alternate scores with those based on total scores. Solutions with few factors showed a difference; Perceptual Speed and Accuracy and Target Identification proportion correct and speed were on separate factors for the alternate scores solution. After about six factors are extracted the solutions are very similar.

Based on these analyses, we conclude the following. The correlations between speed and proportion correct total scores for Perceptual Speed and Accuracy and Target Identification are inflated somewhat due to item interdependence. However, the moderate positive correlation between the scores does not appear to be an artifact of this interdependence and the factor structure is not substantially altered when scores from alternate halves are used in place of total scores. Therefore, we decided to compute separate composites for the speed and proportion correct scores as described above.

Comparison of Longitudinal Initial Sample and Sample 2

To verify the composite scoring of the computer-administered tests, we compared factor analyses on the LV Initial Sample and the LV Sample 2. A principal factor analysis of computer test scores, ASVAB scores, and spatial test scores using Longitudinal Initial Sample data appears in Table 3.46. The 10-factor solution shown was selected after review of several solutions. The corresponding analysis for Longitudinal Sample 2 data appears in Table 3.47.

Table 3.46

Principal Factor Analysis of Total Scores for 33 Cognitive Predictors: Longitudinal Initial Sample

Label	1	2	3	4	5	6	7	8	9	10	R ² *
Assembling Objects	.67*	.21	-.15	.08	.09	.04	-.05	.01	.05	-.12	.55
Reasoning Test	.61*	.24	-.12	.03	.23	.08	-.07	-.06	.06	.07	.52
Maze Test	.59*	.11	-.31	-.03	-.01	.19	-.08	.03	.00	.03	.50
Object Rotation Test	.59*	.10	-.18	-.03	.01	.14	-.05	.08	.05	-.02	.42
Orientation Test	.58*	.23	-.19	.06	.20	-.05	.00	.07	.04	.00	.47
Map Test	.55*	.38	-.15	.05	.26	.04	-.04	.07	.06	.05	.55
ASVAB General Science	.20	.79*	-.12	-.02	.13	-.07	.01	.04	.03	.00	.70
ASVAB Word Knowledge	.16	.77*	-.06	.03	.07	-.05	.02	-.05	.05	.02	.64
ASVAB Electronics Inf	.24	.62*	-.15	.01	.08	-.09	.05	.37	.04	-.01	.61
ASVAB Paragraph Comp	.15	.61*	-.03	.06	.11	.13	-.03	-.05	.06	.04	.44
ASVAB Mechanical Comp	.46*	.51*	-.27	.01	.21	-.08	.03	.30	.04	-.01	.68
Target Tracking 1	-.19	-.13	.82*	.02	-.04	-.01	.09	-.03	-.05	.04	.74
Target Tracking 2	-.24	-.17	.79*	.04	-.05	.00	.07	-.06	-.01	.05	.72
Target Shoot (Dist)	-.08	-.05	.58*	-.06	-.03	.00	.10	.03	-.02	-.04	.37
Cannon Shoot	-.20	-.04	.52*	-.01	-.09	-.03	.08	-.04	-.02	-.08	.33
Movement Time Pooled	-.15	-.02	.31*	.11	.02	-.09	.09	-.09	.19	-.19	.23
Perceptual Spd/Acc DT	-.12	-.02	.12	.69*	-.02	-.14	.25	.07	-.02	-.16	.61
Perceptual Spd/Acc PC	.10	.04	-.03	.65*	.04	.13	.05	-.06	.15	-.04	.49
Target Ident DT	-.38*	-.20	.27	.51*	.04	-.08	.22	.03	-.07	-.06	.58
Target Ident PC	.12	.09	-.09	.47*	.02	-.02	.04	.02	.04	-.10	.27
ASVAB Arith Reasoning	.36	.40	-.09	.05	.58*	.12	-.01	.10	.08	-.03	.67
ASVAB Math Knowledge	.32	.41*	-.05	.01	.57*	.19	-.07	-.04	.03	.00	.64
Number Memory PC	.15	.10	-.13	.27	.39*	.20	-.06	.00	.07	.14	.34
Number Memory DT	-.15	-.22	.07	.26	-.38*	-.29*	.13	.00	-.09	.03	.40
ASVAB Number Oper	.04	-.10	-.02	-.01	.24	.66*	-.13	-.02	.00	.01	.53
ASVAB Coding Speed	.13	.01	-.02	.02	.02	.61*	-.10	-.04	.03	.06	.41
Choice React Time DT	-.04	.01	.07	.13	-.05	-.12	.61*	.05	-.03	.00	.42
Simple React Time DT	-.02	.02	.13	.05	-.02	-.05	.51*	-.01	.01	.02	.28
Short-Term Memory DT	-.16	.00	.14	.29	-.06	-.12	.43*	-.02	-.05	-.24	.39
ASVAB Auto/Shop	.27	.46*	-.22	.03	.02	-.17	.10	.47*	.03	-.04	.59
Choice React Time PC	.06	.03	.00	.06	.03	.05	.03	-.01	.40*	.05	.18
Simple React Time PC	.01	.03	-.01	.01	.01	-.02	-.04	.02	.33*	-.01	.11
Short-Term Memory PC	.15	.05	-.10	.23*	.06	.14	-.06	-.06	.19	.28*	.23
Eigenvalue	3.26	3.26	2.56	1.71	1.28	1.22	1.05	0.51	0.42	0.27	15.59

Note: N = 6436. Asterisks mark highest loadings in row plus any loading that has squared value greater than 1/2 of the square of the largest loading. Squared multiple correlations were used as initial communality estimates.

*R² = communality (sum of squared factor loadings) for variables.

Table 3.47

Principal Factor Analysis of Total Scores for 33 Cognitive Predictors: Longitudinal Sample 2

Label	ASVBA-Gen 1	Spatial 2	Psychomotor 3	Cpx-SpdAcc 4	Number 5	ASVAB-Spd 6	Basic-Spd 7	Basic-Acc 8	ASVAB-Tec 9	Memory 10	R^2 ^a
ASVAB General Science	.79*	.21	-.13	-.02	.14	-.07	.00	.06	-.03	.01	.71
ASVAB Word Knowledge	.77*	.14	-.03	.03	.10	-.04	.03	.07	-.10	.08	.65
ASVAB Electronics Inf	.65*	.24	-.17	.02	.05	-.10	.06	.03	.26	-.05	.60
ASVAB Paragraph Comp	.61*	.14	-.03	.02	.11	.13	-.04	.07	-.06	.06	.43
ASVAB Mechanical Comp	.53*	.48*	-.24	.02	.17	-.11	.02	.01	.23	-.04	.66
ASVAB Auto/Shop	.51*	.28	-.19	.01	.00	-.17	.09	.03	.39*	-.10	.58
Assembling Objects	.20	.69*	-.15	.06	.10	.04	-.06	.04	.02	.13	.57
Maze Test	.11	.60*	-.27	-.07	-.01	.18	-.07	.04	-.02	-.01	.49
Reasoning Test	.25	.59*	-.12	.03	.24	.04	-.08	.04	-.09	.10	.51
Object Rotation Test	.11	.58*	-.18	-.05	.01	.12	-.03	.03	.06	.00	.40
Orientation Test	.23	.56*	-.18	.05	.20	-.05	-.01	.04	.04	.02	.45
Map Test	.37	.55*	-.14	.03	.26	.06	-.01	.05	.05	.05	.54
Target Tracking 1	-.13	-.19	.81*	.03	-.06	-.01	.08	-.07	-.01	.02	.73
Target Tracking 2	-.19	-.23	.78*	.05	-.05	-.01	.05	-.03	-.06	.05	.71
Target Shoot (Dist)	-.04	-.08	.58*	-.06	-.01	-.01	.08	-.04	.04	-.04	.36
Cannon Shoot	-.06	-.21	.50*	.00	-.08	-.03	.08	-.03	-.04	-.08	.32
Movement Time Pooled	-.02	-.13	.32*	.11	.01	-.06	.12	.25*	-.08	-.15	.24
Perceptual Spd/Acc DT	-.03	-.11	.10	.71*	-.02	-.13	.19	.02	.06	-.12	.61
Perceptual Spd/Acc PC	.05	.11	-.03	.63*	.06	.14	.04	.16	-.02	.09	.47
Target Ident DT	-.20	-.36	.25	.55*	.01	-.05	.15	-.11	.03	-.05	.58
Target Ident PC	.09	.12	-.11	.48*	.01	.01	.03	.00	-.03	.14	.29
ASVAB Arith Reasoning	.41	.38	-.08	.05	.58*	.12	-.01	.04	.06	-.02	.68
ASVAB Math Knowledge	.39	.34	-.04	.02	.57*	.18	-.05	.01	-.06	.03	.63
Number Memory DT	-.21	-.13	.06	.27	-.40*	-.29*	.14	-.07	.03	.03	.41
Number Memory PC	.09	.15	-.13	.26	.37*	.23	-.06	.02	.04	.14	.33
ASVAB Number Oper	-.11	.03	-.02	-.01	.24	.66*	-.10	-.01	-.01	.01	.51
ASVAB Coding Speed	-.01	.11	-.01	.02	.02	.62*	-.12	.03	-.02	.07	.42
Choice React Time DT	.02	-.04	.06	.18	-.05	-.16	.56*	-.10	.04	-.03	.38
Simple React Time DT	.02	-.03	.13	.04	-.02	-.04	.50*	.01	-.01	.02	.27
Short-Term Memory DT	-.01	-.18	.13	.36*	-.06	-.14	.37*	.01	-.01	-.22	.39
Choice React Time PC	.04	.05	-.02	.07	.02	.02	.05	.41*	-.01	.06	.18
Simple React Time PC	.05	.03	-.01	-.02	.00	.01	-.09	.31*	.01	-.01	.11
Short-Term Memory PC	.06	.15	-.11	.21	.06	.13	-.05	.11	-.04	.31*	.21
Eigenvalue	3.37	3.27	2.44	1.83	1.28	1.23	.88	.42	.34	.31	15.41

Note: N = 6435. Asterisks mark highest loadings in row plus any loading that has squared value greater than 1/2 of the square of the largest loading.

^a Squared multiple correlations were used as initial communality estimates.

^b R^2 = communality (sum of squared factor loadings) for variables.

The two solutions are virtually the same. One difference, albeit small, has to do with the amount of variance explained by the Spatial and ASVAB-General factors. In the Initial Sample solution, the two factors explained essentially the same amount of variance. In the Sample 2 solution, ASVAB-General explains slightly more variance than does Spatial.

Both solutions illustrate some of the findings that emerge consistently in analyses of these scores. The spatial test scores group to form one factor, and the psychomotor test scores form a factor. The Number Memory Test scores load with ASVAB Math Knowledge and Arithmetic Reasoning on a Number factor. Again, Perceptual Speed and Accuracy and Target Identification proportion correct and time scores load together on a factor. The Simple and Choice Reaction Time Test scores form separate speed and accuracy factors, and Short-Term Memory proportion correct forms a factor on its own.

Reliability Estimates for Computer Test Composites

Split-half scores for each basic score were used to compute reliability estimates for the eight computer composites (Table 3.48). The reliability estimates for the Movement Time, Psychomotor, Perceptual Speed, and Basic Speed composites are quite high; the mean is .95 in both samples. The estimates for the Short-Term Memory and Number Speed and Accuracy composites are a bit lower (i.e., .80-.83). Recall that these two composites are single-test composites, formed by combining the proportion correct and time scores for the test. The accuracy composites (Basic and Perceptual Accuracy) produced the lowest internal consistency estimates. As described earlier, the computer tests were designed in such a way that examinees would be able to get most items they attempt correct. Thus, the constituent proportion correct basic scores for these composites have little variance and are somewhat less reliable than the other scores.

Table 3.48

Internal Consistency Estimates^a for Computer Test Composites: Longitudinal Initial Sample and Sample 2

Composite	Initial Sample (N = 6565)	Sample 2 (N = 6563)
Movement Time	.97	.97
Psychomotor	.94	.94
Perceptual Speed	.98	.98
Perceptual Accuracy	.75	.77
Basic Speed	.92	.92
Basic Accuracy	.62	.63
Short-Term Memory	.80	.81
Number Speed and Accuracy	.83	.83

^aReliabilities are Spearman-Brown corrected split half.

ANALYSES OF COGNITIVE PREDICTOR COMPOSITE SCORES

Based on the analyses presented above, 13 composites of cognitive test scores were recommended for inclusion in the validity analyses: eight computer test composites, one spatial test composite (see Figure 3.4), and the four ASVAB composites (Verbal, Technical, Quantitative, and Speed) that have been identified in previous research (Kass, Mitchell, Grafton, & Wing, 1982). Correlations between these 13 composite scores are shown in Table 3.49 for both the Initial Sample and Sample 2. Compared across the two samples, the correlations are similar in magnitude and pattern.

ASVAB Verbal, Quantitative, and Technical composites correlate in the .45-.65 range with each other and with the Spatial composite. The Spatial composite is moderately correlated with most of the computer composites, particularly Psychomotor, Perceptual Speed, Number Speed and Accuracy, and Short-Term Memory. Perceptual Speed and Perceptual Accuracy are more highly correlated with each other than with any other composites, and this correlation is negative. That is, highly accurate examinees were also slower responders on the two tests forming these composites. Number Speed and Accuracy is most highly correlated with ASVAB Quantitative. Short-Term Memory has low/moderate (around .30) correlations with several other composites, but no large correlation with any other composite. Movement Time is most highly correlated with the Psychomotor composite, and the two basic composites (Basic Speed and Basic Accuracy) have low correlations with almost all of the other composites.

The means, standard deviations, and effect sizes for composite scores computed by gender are shown in Table 3.50. The largest effect sizes were observed for the Psychomotor, ASVAB Technical, ASVAB Speed, and Movement Time composites. Means for men were over one standard deviation greater than the means for women on the Psychomotor and ASVAB Technical composites. On ASVAB Speed means for women were about two-thirds of a standard deviation greater than the means for men, and on Movement time men outperformed women by about one-half standard deviation.

The means, standard deviations, and effect sizes for composite scores computed by race are shown in Table 3.51. When means for blacks and whites are compared, three composites yield effects greater than one standard deviation difference: ASVAB Verbal, ASVAB Technical, and Spatial. One-half or more standard deviation difference is observed for another three composites: ASVAB Quantitative, Psychomotor, and Number Speed and Accuracy.

When Hispanic and white means are compared, the effect sizes observed for the two samples (Initial and Sample 2) are more variable than those for blacks and whites. The effect sizes are not as stable for the smaller sample of Hispanics. The largest effect sizes, nearing one standard deviation, are for ASVAB Verbal and ASVAB Technical. Most of the other effects were less than one-half standard deviation.

SCORING AND FORMING COMPOSITES FOR THE ABLE INVENTORY

After a brief review of the development and content of the ABLE inventory, this section will cover the following areas: (a) data screening, (b) analysis of the appropriateness of the scoring procedures developed and

Table 3.49

Correlations Between Cognitive Composites*: Longitudinal Initial Sample and Sample 2

	Spatial	ASVAB Verbal	ASVAB Quant	ASVAB Speed	ASVAB Techn	Psychomotor	Perc Speed	Perc Accur	Number Sp/Abc	Basic Speed	Basic Accur	Short-Term Memory	Move Time
<u>Longitudinal Initial Sample (N = 6436)</u>													
Spatial	--												
ASVAB Verbal	.45	--											
ASVAB Quantitative	.58	.54	--										
ASVAB Speed	.18	-.01	.23	--									
ASVAB Technical	.56	.63	.50	-.09	--								
Psychomotor	.44	.23	.25	.07	.38	--							
Perceptual Speed	.33	.17	.17	.16	.16	.30	--						
Perceptual Accuracy	.20	.13	.14	.07	.12	.11	-.43	--					
Number Speed													
Accuracy	.38	.30	.57	.33	.24	.23	.19	.17	--				
Basic Speed	.11	.01	.08	.18	-.04	.18	.30	-.09	.15	--			
Basic Accuracy	.09	.07	.09	.03	.07	.04	.02	.10	.09	.01	--		
Short-Term Memory	.29	.10	.19	.21	.09	.23	.32	.08	.25	.28	.10	--	
Movement Time	.22	.07	.08	.10	.15	.31	.23	-.06	.10	.11	-.09	.19	--
<u>Longitudinal Sample 2 (N = 6435)</u>													
Spatial	--												
ASVAB Verbal	.45	--											
ASVAB Quantitative	.59	.54	--										
ASVAB Speed	.15	-.02	.22	--									
ASVAB Technical	.56	.64	.48	-.12	--								
Psychomotor	.43	.23	.24	.06	.37	--							
Perceptual Speed	.33	.19	.17	.13	.17	.28	--						
Perceptual Accuracy	.19	.14	.16	.09	.11	.12	-.44	--					
Number Speed													
Accuracy	.37	.29	.56	.34	.21	.23	.19	.12	--				
Basic Speed	.11	.01	.07	.19	-.04	.17	.27	-.08	.17	--			
Basic Accuracy	.10	.10	.08	.03	.07	.05	.04	.08	.07	.04	--		
Short-Term Memory	.30	.13	.21	.21	.08	.23	.31	.07	.25	.26	.08	--	
Movement Time	.19	.05	.08	.08	.13	.31	.21	-.05	.11	.11	-.10	.20	--

*Before composites were formed, T-scores were computed for all basic scores, and all distance and time basic scores were reflected such that a high score indicates "better" performance.

Table 3.50

Cognitive Composite Score Means and Effect Sizes by Gender for Longitudinal Initial Sample and Sample 2

Composite	Male			Female			Effect Size ^a (d)
	N	Mean	SD	N	Mean	SD	
ASVAB Verbal							
Longitudinal Initial	6037	157.89	16.93	890	155.27	14.96	.15
Longitudinal Sample 2	6042	158.17	17.01	884	155.36	15.03	.16
ASVAB Quantitative							
Longitudinal Initial	6037	103.62	13.66	890	101.54	11.88	.15
Longitudinal Sample 2	6042	103.44	13.85	884	102.05	12.06	.10
ASVAB Speed							
Longitudinal Initial	6037	106.84	11.38	890	113.96	10.64	-.63
Longitudinal Sample 2	6042	106.64	11.36	884	113.79	10.78	-.63
ASVAB Technical							
Longitudinal Initial	6037	159.11	21.37	890	137.03	16.30	1.06
Longitudinal Sample 2	6042	159.47	20.92	884	136.71	15.77	1.11
Spatial							
Longitudinal Initial	6038	301.68	45.32	891	289.22	42.12	.27
Longitudinal Sample 2	6047	301.81	44.72	885	288.25	42.75	.30
Psychomotor							
Longitudinal Initial	5928	205.99	28.17	813	165.77	29.07	1.42
Longitudinal Sample 2	5912	205.99	27.78	809	165.80	30.45	1.43
Perceptual Speed							
Longitudinal Initial	6007	100.82	17.38	886	94.28	17.21	.37
Longitudinal Sample 2	5995	100.83	17.41	887	94.29	17.20	.37
Perceptual Accuracy							
Longitudinal Initial	6013	99.64	16.37	888	103.16	14.77	-.21
Longitudinal Sample 2	6000	99.60	16.42	887	103.27	15.34	-.22
Number Speed & Accuracy							
Longitudinal Initial	6013	100.37	15.39	890	97.49	14.82	.18
Longitudinal Sample 2	6018	100.29	15.39	885	98.16	15.25	.13
Basic Speed							
Longitudinal Initial	6065	100.02	16.88	894	100.11	14.33	-.01
Longitudinal Sample 2	6071	100.06	16.49	889	99.67	17.97	.02
Basic Accuracy							
Longitudinal Initial	6065	99.83	15.43	894	101.86	13.73	-.13
Longitudinal Sample 2	6071	99.81	15.56	889	101.65	14.80	-.11
Short-term Memory							
Longitudinal Initial	6038	99.89	15.21	888	100.77	13.57	-.05
Longitudinal Sample 2	6033	99.77	15.01	890	101.53	13.48	-.11
Movement Time							
Longitudinal Initial	5936	50.65	9.90	879	45.59	9.52	.51
Longitudinal Sample 2	5927	50.72	9.80	878	45.12	9.99	.57

^a d is the standardized mean difference between males and females. A negative sign indicates superior performance by females.

Table 3.51

Cognitive Composite Score Means and Effect Sizes by Race for Longitudinal Initial Sample and Sample 2

Composite	White			Black			Hispanic			Other		
	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
ASVAB Verbal												
Longitudinal Initial	4749	162.50	14.59	1705	145.75	15.44	1.13	237	148.57	16.24	233	151.77
Longitudinal Sample 2	4756	162.70	14.61	1699	145.93	15.53	1.12	233	151.88	16.84	233	150.29
ASVAB Quantitative												
Longitudinal Initial	4749	106.06	13.29	1705	95.89	11.20	.79	237	101.27	12.02	233	104.66
Longitudinal Sample 2	4756	106.00	13.34	1699	95.76	11.64	.79	233	101.95	12.82	233	103.09
ASVAB Speed												
Longitudinal Initial	4749	107.52	11.55	1705	107.89	11.55	-.03	237	108.88	11.42	233	110.26
Longitudinal Sample 2	4756	107.18	11.56	1699	108.22	11.45	-.09	233	108.71	11.19	233	108.91
ASVAB Technical												
Longitudinal Initial	4749	164.01	19.42	1705	136.93	16.37	1.45	237	146.72	18.47	233	149.38
Longitudinal Sample 2	4756	164.00	19.11	1699	137.80	16.39	1.42	233	147.94	18.90	233	149.44
Spatial												
Longitudinal Initial	4741	313.29	40.96	1707	264.34	37.11	1.22	237	292.41	40.09	233	299.90
Longitudinal Sample 2	4740	312.38	40.57	1710	266.96	38.53	1.13	233	294.11	44.71	233	298.80
Psychomotor												
Longitudinal Initial	4673	206.52	28.99	1598	185.33	32.20	.71	230	201.14	31.86	229	201.19
Longitudinal Sample 2	4676	206.03	28.97	1583	186.67	32.14	.64	225	201.14	30.04	221	201.01
Perceptual Speed												
Longitudinal Initial	4737	101.73	16.66	1677	94.84	18.74	.40	236	99.75	16.82	232	101.03
Longitudinal Sample 2	4731	101.75	16.58	1677	95.01	19.29	.38	228	100.81	16.35	230	99.22
Perceptual Accuracy												
Longitudinal Initial	4744	101.08	15.48	1678	97.43	17.67	.22	236	100.50	16.89	232	99.02
Longitudinal Sample 2	4735	100.68	16.02	1678	98.08	17.29	.15	228	100.52	15.25	230	101.56
Number Speed & Accuracy												
Longitudinal Initial	4748	102.14	14.56	1677	94.53	15.98	.50	237	96.17	16.08	230	99.53
Longitudinal Sample 2	4744	102.04	14.43	1677	94.92	16.49	.47	233	97.57	15.47	233	97.94
Basic Speed												
Longitudinal Initial	4772	99.87	15.60	1700	100.35	19.34	-.02	239	101.29	14.43	237	99.67
Longitudinal Sample 2	4768	100.04	15.37	1709	99.75	19.28	.01	234	101.58	13.68	233	99.59
Basic Accuracy												
Longitudinal Initial	4772	100.94	13.53	1700	98.04	18.82	.19	239	100.54	13.77	237	97.04
Longitudinal Sample 2	4768	101.11	13.78	1709	97.79	18.61	.21	234	98.08	17.40	233	96.34
Short Term Memory												
Longitudinal Initial	4761	100.69	14.11	1686	97.82	17.29	.19	236	101.44	14.13	232	100.57
Longitudinal Sample 2	4763	100.76	14.19	1684	97.55	16.84	.21	230	100.91	12.46	230	101.43
Movement Time												
Longitudinal Initial	4706	50.77	9.48	1639	47.62	11.06	.31	232	49.99	10.13	227	51.17
Longitudinal Sample 2	4695	50.50	9.75	1644	47.92	10.46	.26	224	51.69	9.81	226	53.10

^ad is the standardized mean difference between two subgroups' scores. All effect sizes are relevant to the white group. A negative value indicates superior performance by the comparison group (black, Hispanic, or other).

refined in earlier stages of the research, (c) descriptive statistics for the Initial Sample, (d) subgroup differences, (e) uniqueness analysis, and (f) composite formation.

Development and Content of the ABLE Inventory

The ABLE (Assessment of Background and Life Experiences) was developed to measure biodata and temperament constructs. Like the computer and other paper-and-pencil predictors, this inventory has been developed iteratively over several years. The constructs measured were selected as the result of an extensive literature review, which is reported in Hough (1986). The development and pilot testing of the ABLE are described in Hough, Barge, and Kamp (1987); the field testing is described in Hough, McGue, Houston, and Pulakos (1987); and the evaluation of the Trial Battery and Revised Trial Battery versions of the ABLE is reported in Hough, McCloy, Ashworth, and Hough (1987).

Several changes were made to the Trial Battery version of the ABLE to form the Experimental Battery version. These changes were made in two phases. First, 10 items were deleted because they were part of the AVOICE inventory (see following section), had low item-total correlations, were difficult to interpret, or appeared inappropriate for the age group. The inventory that resulted is called the Revised Trial Battery version of the ABLE. Then 16 items were modified based on item statistics and reviewers' comments, and the instructions were changed slightly to allow for the use of a separate answer sheet. This most recent version is the Experimental Battery version.

The Revised Trial Battery ABLE is simply a way of scoring the CV data; the Experimental Battery ABLE was administered to the Longitudinal Validation (LV) sample. The distinction between these versions of the inventory is important because the results obtained for the Experimental Battery (LV sample) will be compared to those obtained for the Revised Trial Battery (CV sample).

Data Screening

Two methods were employed to screen, from the Initial Sample, respondents who appear to have been either unwilling to attend to the inventory or unable to comprehend the questions. We used the same procedures that were used to screen the CV sample: records were removed from the data set (a) if respondents answered fewer than 90 percent of the questions, or (b) if they answered incorrectly three or more of the eight questions in the Non-Random Response scale, which includes questions that should be answered correctly by all persons who carefully read and respond to the questions. Endorsement rates appear in Table 3.52 for the eight items in this scale, revealing that most people indeed answered each question correctly.

The number and percentage of persons eliminated from the CV and LV samples by the missing data and Non-Random Response scale screens are shown in Table 3.53. In comparison to the CV sample, more persons were screened from the LV sample by the Non-Random Response scale screen and fewer were removed by the missing data screen. This resulted in a slightly smaller proportion of persons screened by the two procedures combined. In general, a high rate

Table 3.52**Longitudinal Validation: Option Endorsement Rates for Items on the ABLE Non-Random Response Scale**

Item	Correct Response Option	N	Option 1 (%)	Option 2 (%)	Option 3 (%)
NR1	3	6981	2.7	2.4	<u>94.9</u>
NR2	1	6981	<u>93.6</u>	2.9	3.5
NR3	2	6964	8.4	<u>89.5</u>	2.0
NR4	1	6967	<u>92.7</u>	5.1	2.3
NR5	2	6962	4.4	<u>93.1</u>	2.5
NR6	1	6964	<u>91.8</u>	5.3	2.8
NR7	3	6956	4.9	7.0	<u>88.1</u>
NR8	2	6948	3.3	<u>95.3</u>	1.4

Table 3.53**Comparison of CV and LV ABLE Data Screening Results**

	Number		Percent	
	CV	LV	CV	LV
Number of Inventories Scanned	9359	7000	100.0	100.0
Deleted Using Overall Missing Data Screen (Decision Rule: If missing data greater than 10%, delete inventory)	171	40	1.8	0.6
Deleted Using Non-Random Response Scale Screen (Decision Rule: If fewer than 6 of 8 responses are "correct," delete inventory)	684	565	7.3	8.1
Respondents Passing Screening Criteria	8504	6395	90.9	91.3

(over 90 percent) of the sample appeared to read and answer the questions carefully.

For inventories surviving these screens, missing data were treated in the following way. If more than 10 percent of the item responses in a scale were missing, the scale score was not computed; instead the scale score was treated as missing. If there were missing item responses for a scale but the percent missing was equal to or less than 10 percent, then the person's average item response score for that scale was computed and used for the missing response.

Analyses to Verify Appropriateness of the Scoring Procedures

Scale scores were formed according to the scoring procedure developed during earlier phases of the project. (See Hough, Barge, & Kamp, 1987; Hough, McCloy, Ashworth, & Hough, 1987; Hough, McGue, Houston, & Pulakos, 1987 for a complete description of the analyses and methods used.) Each item was then correlated with each ABLE scale, and these correlations were compared to within-scale item-total correlations (with the item removed from the total). Few items correlated substantially higher with another scale than with their own scale. In total, 13 of the 199 items (7%) correlated higher with another scale than with their own scale by a margin of .05 or more.

Given this relatively small number, we decided to retain the rationally and empirically developed scoring procedure established in the previous research phases. Although the scales could be made more internally consistent, they were not intended to be highly homogeneous. ABLE scales cut across both the temperament and biodata domains and measure fairly broad constructs. Our aim was to maintain the conceptual framework established previously while maximizing the external (i.e., predictive) validity of the scales.

Comparison of Descriptive Statistics for the Revised Trial Battery and Experimental Battery

In the CV research phase, the ABLE Revised Trial Battery was found to have adequate reliability and stability and to correlate with job performance in the Army (McHenry, et al., 1990). Therefore, the Revised Trial Battery descriptive statistics were used as a benchmark against which to compare the psychometric characteristics of the Experimental Battery version of the ABLE. The means, standard deviations, and internal consistency reliabilities for each of 15 ABLE scale scores for the Revised Trial Battery and Experimental Battery are reported in Table 3.54. Test-retest reliabilities obtained for the Revised Trial Battery are also presented; test-retest data were not collected during Longitudinal Validation.

Several interesting findings are revealed in this table. First, LV respondents tended to score higher than CV respondents. In particular, LV respondents had higher mean scores on the Cooperativeness, Nondelinquency, Traditional Values, and Internal Control scales, on which they scored more than half a standard deviation higher than CV respondents. This probably can be attributed to differences between the LV and CV testing conditions: LV respondents completed the inventory in the first few days on the job, in contrast to CV respondents, who completed the inventory after a year or two in the Army. LV respondents may have believed (in spite of being told the contrary by the test administrators) that their responses to the inventory would affect their career in the Army and thus responded in a more favorable direction. Indeed, LV respondents on average scored more than one third of a standard deviation higher than CV respondents on the Unlikely Virtues scale, a measure of the tendency to respond in a socially desirable manner. Also, for those scales showing the greatest increase in scores, there was a decrease in the standard deviation. Internal consistency reliabilities remained acceptable for these scales, however, with reliabilities for the 11 content scales ranging from .64 (Traditional Values) to .86 (Work Orientation).

Table 3.54

Comparison of ABLE Scale Scores and Reliabilities From the Revised Trial (CV) and Experimental (LV) Batteries

ABLE Scale	No. of Items	Sample Size		Mean		SD		Effect Size ^b		Median Item-Total Correlation ^c		Internal Consistency Reliability (Alpha)		Test-Retest Reliability ^d
		CV	LV ^a	CV	LV	CV	LV	CV	LV	CV	LV	CV	LV	
Emotional Stability	17	8522	6395	39.0	40.0	5.45	5.61	.18	.45	.39	.45	.81	.84	.74
Self-Esteem	12	8472	6394	28.4	28.7	3.70	3.96	.08	.44	.39	.44	.74	.78	.78
Cooperativeness	18	8494	6395	41.9	44.4	5.28	4.94	.49	.39	.39	.39	.81	.80	.76
Conscientiousness	15	8504	6385	35.1	36.7	4.31	4.10	.38	.34	.34	.34	.72	.73	.74
Nondevinquency	20	8482	6390	44.2	47.8	5.91	5.52	.63	.35	.36	.35	.81	.78	.80
Traditional Values	11	8461	6387	26.6	29.0	3.72	2.94	.70	.30	.36	.30	.69	.64	.74
Work Orientation	19	8498	6394	42.9	45.2	6.06	6.07	.38	.41	.41	.42	.84	.86	.78
Internal Control	16	8485	6393	38.0	41.7	5.11	4.38	.77	.37	.39	.37	.78	.76	.69
Energy Level	21	8488	6390	48.4	50.4	5.97	5.99	.33	.42	.38	.42	.82	.84	.78
Dominance	12	8477	6392	27.0	27.2	4.28	4.65	.05	.44	.44	.51	.80	.84	.79
Physical Condition	6	8500	6395	14.0	13.3	3.04	3.01	-.23	.58	.60	.58	.84	.81	.85
Unlikely Virtues	11	8511	6393	15.5	16.8	3.04	3.38	.41	.34	.34	.35	.63	.66	.63
Self-Knowledge	11	8508	6392	25.4	26.2	3.33	3.12	.35	.23	.36	.23	.65	.59	.64
Non-Random Response	8	8559	6395	7.7	7.7	.59	.58	.00	--	--	--	--	--	.30
Poor Impression	23	8492	6393	1.5	1.2	1.85	1.65	-.17	.21	.20	.21	.63	.62	.61

^aInitial longitudinal sample screened for missing data and random responding.

^bEffect Size = $(\text{Mean}_{LV} - \text{Mean}_{CV}) / \text{Pooled Standard Deviation}$. Positive effect sizes indicate higher mean scores in the LV sample.

^cEach item was correlated with the sum of the remaining items in the scale.

^dn = 408-414.

Analysis of Subgroup Differences

ABLE means and standard deviations by gender are presented in Table 3.55. Compared to men, women appeared less delinquent (they scored slightly higher on Nondelinquency), they had higher internal control, and they knew themselves better. Men tended to appear in better physical condition, scoring on average a half a standard deviation higher on the Physical Condition scale. In general, however, gender differences tended to be small.

Table 3.55

Longitudinal Validation: ABLE Scale Score Means and Effect Sizes by Gender

ABLE Scale	Male		Female		Effect Size ^a (d)
	Mean	SD	Mean	SD	
Emotional Stability	40.1	5.60	39.1	5.64	.18
Self-Esteem	28.8	3.96	28.4	3.97	.09
Cooperativeness	44.4	4.93	44.9	4.93	-.11
Conscientiousness	36.6	4.11	37.3	3.95	-.18
Nondelinquency	47.5	5.49	49.3	5.44	-.33
Traditional Values	28.9	2.97	29.3	2.74	-.11
Work Orientation	45.1	6.08	46.0	5.96	-.14
Internal Control	41.6	4.44	42.6	3.84	-.25
Energy Level	50.4	6.02	50.7	5.82	-.04
Dominance	27.3	4.60	26.5	4.89	.17
Physical Condition	13.5	2.94	1.9	3.07	.54
Unlikely Virtues	16.8	3.36	16.8	3.51	.00
Self-Knowledge	26.2	3.12	26.8	3.08	-.21
Non-Random Response	7.7	.59	7.7	.54	-.05
Poor Impression	1.2	1.64	1.3	1.73	-.07

Note. N for males = 5519-5529; N for females = 865-866.

^a d is the standardized mean difference between males' and females' scores. A positive value indicates higher scores for males; a negative value indicates higher scores for females.

ABLE means and standard deviations by race are presented in Table 3.56. Race differences tended to be quite small as well, with blacks scoring slightly higher than whites, Hispanics, or other groups on seven of the 11 content scales. Using the ABLE as a selection device will not adversely impact women or minorities.

Uniqueness Analyses

It is important that the ABLE scales tap reliable variance that is unique from the ASVAB. If so, these scales have the potential to improve the prediction of job performance above and beyond prediction by the ASVAB.

Longitudinal Validation: ABLE Scale Means and Effect Sizes by Race

d is the standardized mean difference between two subgroup scores. All effect sizes in this table are relative to the white subgroup. A positive effect size indicates that whites score higher than the minority, and a negative value indicates that whites score lower.

To investigate this question, the 10 ASVAB subtests were entered into regressions to predict each of the ABLE scales. The amount of variance shared by an ABLE scale and the ASVAB scales (R^2) subtracted from the reliable variance measured by the scale (the scale's reliability) can be interpreted as the amount of reliable variance independent of the ASVAB (ASVAB Uniqueness).

Table 3.57 presents these squared multiple regression coefficients along with reliability coefficients and uniqueness estimates for ABLE scale scores. As shown in this table, the ASVAB accounts for only 1 to 3 percent of the variance in the ABLE scale scores. Thus, the uniquenesses are quite high, and the ABLE has good potential for contributing to the prediction of job performance.

Table 3.57

Comparison of Reliability Coefficients, Multiple Regression Coefficients,^a and Uniqueness Estimates for ABLE Scale Scores: CV/LV

ABLE Scale	No. Items	Alpha Coefficient ^b	ASVAB R-Squared ^c	Uniqueness ^d
Emotional Stability	17	.81/.84	.02/.02	.79/.82
Self-Esteem	12	.74/.78	.04/.03	.70/.75
Cooperativeness	18	.81/.80	.01/.01	.80/.79
Conscientiousness	15	.72/.73	.02/.01	.70/.72
Nondelinquency	20	.81/.78	.03/.02	.78/.76
Traditional Values	11	.69/.64	.01/.00	.68/.64
Work Orientation	19	.84/.86	.01/.02	.83/.84
Internal Control	16	.78/.76	.02/.02	.76/.74
Energy Level	21	.82/.84	.01/.01	.81/.83
Dominance	12	.80/.84	.01/.02	.79/.82
Physical Condition	6	.84/.81	.01/.01	.83/.80
Unlikely Virtues	11	.63/.66	.09/.04	.54/.62
Self-Knowledge	11	.65/.59	.02/.01	.63/.58
Non-Random Response	8	---	.06/.02	---
Poor Impression	23	.63/.62	.01/.01	.62/.61

^aVersus all ASVAB subtests.

^bN = 8064/6385.

^c R^2 is adjusted for shrinkage (i.e., cross-validity estimated); N = 7091/4930.

^dAlpha reliability minus ASVAB adjusted R^2 .

Formation of ABLE Composites

As mentioned previously, it is important to attempt to reduce the number of predictors to a more manageable number before entering them into regressions for the prediction of job performance. Each ABLE scale was designed to measure a unique construct, so many of these scales may not fit into clusters with other ABLE scales. Nevertheless, several approaches were taken to identify clusters of scales that might be combined to form meaningful and coherent composite scores.

The rationale for performing both principal components and principal factor analyses was a desire to be thorough. Both types of analyses were performed to provide a wider latitude for selecting a set of factors that rationally were interpretable. While principal factor analysis would produce a set of factors that are more likely to be stable over different samples, principal components analysis was also conducted to discover whether smaller, additional factors that were interpretable might emerge.

The correlations among scale scores for the CV and LV samples are shown in Table 3.58. Interestingly, the correlations between content scale scores were higher in the LV sample for 51 of the 54 intercorrelations. Also, all of the correlations between the Unlikely Virtues scale and the content scales were higher in the LV sample than in the CV sample, indicating perhaps more social desirability bias in LV responses. Greater social desirability bias could account for higher correlations among all of the scales. It also can make it more difficult to identify clusters of scale scores.

Principal components analyses were run on the 11 ABLE content scale scores and the eigenvalues obtained were compared to parallel analysis estimates of eigenvalues for random data (Humphreys & Montanelli, 1975; Montanelli & Humphreys, 1976). The parallel analysis suggested that one component should be retained.

Table 3.59 presents the two-component orthogonal varimax rotation solution. The first component appears to be composed of six scales: Self-Esteem, Dominance, Energy Level, Emotional Stability, Physical Condition, and Work Orientation. The second component is composed of the remaining five content scale scores. Clearly, however, a reduction of the 11 scores into the two scores would serve to hide potentially important clusters of scales.

Principal factor analysis was also performed on the matrix of correlations among the ABLE content scale scores. Parallel analysis was conducted to determine the number of factors whose eigenvalues exceed the levels expected by chance. As many as three factors obtained from the factor analysis have eigenvalues greater than chance levels. Thus, a three-factor principal factor analysis was conducted and is reported in Table 3.60. Unfortunately, however, the third factor is difficult to interpret because no scales load more highly on this factor than on the first two factors.

An effort was then made to replicate the factor analysis results obtained on the CV data. Principal components analysis was performed on all 15 of the ABLE scales and a seven-component solution was selected to mirror the solution selected in the analysis of the CV data. Table 3.61 presents, side-by-side, the factor pattern matrixes for these solutions and reveals an extremely close match. All of the content scales except for Physical Condition clearly load on one of the first three components, and Self-Knowledge, Unlikely Virtues, Physical Condition, and Non-Random Response appear to define their own composites.

Next, the various composite formation models suggested by these factor and component analyses were compared using LISREL (Joreskog & Sorbom, 1986). These models, along with the corresponding degrees of freedom, chi-square, goodness-of-fit, adjusted goodness-of-fit, and root mean square residual, are shown in Table 3.62. Of these models, the second model appears to fit the data best. The Dominance scale was later removed from the first composite

Table 3.58

Comparison of ABLE Scale Intercorrelations From the Revised Trial (CV) and Experimental (LV) Batteries: CV/LV

ABLE Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Emotional Stability														
2. Self-Esteem	.54/.65													
3. Cooperativeness	.50/.54	.44/.51												
4. Conscientiousness	.31/.44	.47/.52	.47/.54											
5. Nondefensiveness	.29/.35	.29/.35	.52/.56	.59/.58										
6. Traditional Values	.22/.27	.27/.29	.43/.45	.56/.52	.62/.58									
7. Work Orientation	.38/.53	.60/.66	.48/.54	.65/.70	.41/.46	.44/.43								
8. Internal Control	.40/.46	.39/.46	.42/.46	.46/.50	.37/.41	.45/.46	.48/.50							
9. Energy Level	.60/.70	.65/.70	.52/.57	.55/.61	.38/.42	.40/.40	.73/.75	.51/.55						
10. Dominance	.42/.54	.64/.67	.33/.40	.40/.44	.18/.22	.22/.23	.52/.57	.29/.34	.54/.58					
11. Physical Condition	.24/.37	.31/.43	.18/.23	.18/.28	.05/.15	.07/.14	.29/.40	.12/.19	.35/.49	.31/.36				
12. Unlikely Virtues	.13/.27	.15/.26	.25/.35	.30/.41	.35/.42	.25/.28	.31/.42	.08/.17	.26/.35	.13/.22	.08/.16			
13. Self-Knowledge	-.02/.05	.26/.24	.18/.21	.30/.25	.15/.13	.20/.14	.29/.26	.18/.18	.22/.21	.26/.23	.14/.13	.02/.01		
14. Non-Random Response	.10/.09	.07/.08	.12/.13	.10/.11	.08/.15	.09/.14	.06/.07	.18/.22	.09/.11	.05/.04	.01/.02	-.24/-.12	.08/.10	
15. Poor Impression	-.58/-.60	-.34/-.43	-.46/-.43	-.34/-.38	-.40/-.37	-.33/-.31	-.30/-.39	-.41/-.42	-.49/-.55	-.24/-.32	-.17/-.29	-.08/-.14	-.00/-.04	.15/-.18

Note. N = CV 8,437-8,522 / LV 6382-6395.

Table 3.59**Principal Components Analysis^a of the ABLE Content Scales (Initial Longitudinal Sample)**

ABLE Scale	Factor 1	Factor 2	h^2 ^b
Self-Esteem	.81	.31	.75
Dominance	.78	.17	.63
Energy Level	.77	.45	.79
Emotional Stability	.73	.32	.63
Physical Condition	.68	-.01	.46
Work Orientation	.66	.53	.71
Nondelinquency	.10	.83	.70
Traditional Values	.04	.82	.68
Conscientiousness	.42	.70	.67
Cooperativeness	.40	.66	.59
Internal Control	.35	.61	.50
Eigenvalue	3.76	3.37	7.11

Note. N = 6200.

^aVarimax rotation.^b h^2 = communality (sum of squared factor loadings) for variables.

Table 3.60**Principal Factor Analysis^a of the ABLE Content Scales (Initial Longitudinal Sample)**

ABLE Scale	Factor 1	Factor 2	Factor 3	h^2 ^b
Self-Esteem	.77	.30	.11	.70
Energy Level	.76	.44	.08	.78
Dominance	.70	.19	.05	.53
Work Orientation	.67	.52	-.15	.74
Emotional Stability	.66	.31	.32	.64
Physical Condition	.53	.09	-.04	.29
Nondelinquency	.15	.74	.06	.57
Traditional Values	.12	.70	.00	.50
Conscientiousness	.44	.67	-.13	.66
Cooperativeness	.39	.59	.21	.55
Internal Control	.36	.53	.13	.43
Eigenvalue	3.30	2.86	.22	6.38

Note. N = 6200.

^aVarimax rotation, 3-factor solution.

^b h^2 = communality (sum of squared factor loadings) for variables.

Table 3.61

Comparison of ABLE Principal Components Analysis^a Results: CV/LV

ABLE Scale	Achievement Orientation	Dependability	Adjustment	Unlikely Virtues	Self-Knowledge	Physical Condition	Non-Random Response	h^2 ^b
Dominance	<u>.81/.85</u>	.03/.06	.13/.10	.04/.07	.15/.10	.09/.08	.03/.00	.71/.76
Self-Esteem	<u>.80/.78</u>	.12/.15	.28/.28	.05/.12	.14/.11	.07/.17	.03/.02	.76/.77
Work Orientation	<u>.70/.67</u>	.50/.40	.05/.11	.16/.30	.06/.13	.13/.21	.01/.01	.79/.78
Energy Level	<u>.68/.66</u>	.36/.31	.36/.39	.06/.18	-.01/.06	.19/.29	.00/.04	.77/.80
Traditional Values	<u>.08/.08</u>	<u>.82/.85</u>	.16/.11	.13/.15	.09/.04	.02/.06	.04/.03	.73/.78
Conscientiousness	.42/.45	<u>.67/.58</u>	.11/.11	.23/.35	.16/.16	.05/.12	.09/.05	.72/.72
Model Iniquity	.01/.04	<u>.67/.61</u>	.37/.32	.40/.50	.14/.11	-.03/.03	.09/.11	.77/.76
Internal Control	.36/.45	<u>.66/.66</u>	.26/.23	-.35/-.15	-.06/-.01	-.02/-.05	-.03/.13	.77/.73
Emotional Stability	.48/.60	.04/.10	<u>.74/.65</u>	.00/.12	-.14/-.08	.06/.09	.02/.02	.80/.81
Cooperativeness	.25/.34	.33/.36	<u>.63/.50</u>	.26/.37	.22/.23	.02/-.06	.09/.07	.70/.69
Poor Impression	-.12/-.21	-.28/-.22	<u>-.81/-.84</u>	.08/.02	.07/.03	-.09/-.16	-.04/-.08	.76/.82
Unlikely Virtues	.16/.20	.21/.14	.02/.02	<u>.83/.88</u>	-.09/-.07	.02/.05	-.21/-.10	.81/.85
Self-Knowledge	.21/.16	.16/.09	-.06/-.02	-.07/-.03	<u>.92/.96</u>	.05/.05	.01/.04	.93/.96
Physical Condition	.23/.28	.02/.05	.10/.15	.02/.05	.05/.05	<u>.96/.92</u>	.00/.00	.99/.96
Non-Random Response	.04/.03	.09/.10	.08/.08	-.16/-.05	.01/.04	.00/.00	<u>.97/.99</u>	.96/1.0
Eigenvalue	2.99/3.29	2.67/2.38	2.09/1.84	1.18/1.49	1.03/1.07	1.01/1.07	1.02/1.04	11.99/12.19

Note: N = 8,348/6,368

^aVarimax rotation.^b h^2 = communality (sum of squared factor loadings) for variables.

Table 3.62

Longitudinal Validation: Comparison of Four ABLE Composite Formation Models, Using LISREL

	MODEL 1	MODEL 2	MODEL 3	MODEL 4
Degrees of Freedom	40	37	38	42
Chi-Square	2079	1843	1962	2083
Goodness-of-Fit Index	.920	.929	.924	.920
Adjusted Goodness-of-Fit Index	.868	.873	.869	.874
Root Mean Square Residual	.080	.072	.077	.076
Composite Formation Model:	Dominance Self-Esteem Work Orientation Energy Level Traditional Values Conscientiousness Monell frequency Internal Control Emotional Stability Cooperativeness Physical Condition	Self-Esteem Work Orientation Energy Level Dominance Traditional Values Conscientiousness Monell frequency Emotional Stability Cooperativeness Internal Control Physical Condition	Self-Esteem Work Orientation Energy Level Dominance Traditional Values Conscientiousness Monell frequency Emotional Stability Cooperativeness Internal Control Physical Condition	Self-Esteem Work Orientation Energy Level Dominance Emotional Stability Traditional Values Conscientiousness Monell frequency Cooperativeness Internal Control Physical Condition

because of its unique potential for the prediction of leadership in the Army. The proposed LV composite model is shown in Figure 3.5, along with the composite model used in the analyses.

Longitudinal Validation Composites

Concurrent Validation Composites

Achievement Orientation

Self-Esteem
Work Orientation
Energy Level

Leadership Potential

Dominance

Dependability

Traditional Values
Conscientiousness
Nondelinquency

Adjustment

Emotional Stability

Cooperativeness

Cooperativeness

Internal Control

Internal Control

Physical Condition

Physical Condition

Achievement Orientation

Self-Esteem
Work Orientation
Energy Level

Dependability

Conscientiousness
Non-Delinquency

Adjustment

Emotional Stability

Physical Condition

Physical Condition

Note. Four ABLE scales were not used in computing CV composite scores. These were Dominance, Traditional Values, Cooperativeness, and Internal Control.

Figure 3.5. Comparison of ABLE composites for the Longitudinal and Concurrent Validations.

We computed split-half scores for each ABLE scale and used them to form split-half composite scores. Table 3.63 presents the split-half reliabilities for the ABLE composite scores, along with ABLE composite score Uniqueness estimates. As expected, these composite scores measure variance that is reliable and unique from the ASVAB.

Table 3.63

Longitudinal Validation: ABLE Composite Score Reliability and Uniqueness Estimates

Composite	Split-Half Reliability ^a	ASVAB R-Squared ^b	Uniqueness ^c
Achievement Orientation	.91	.02	.89
Leadership Potential	.83	.02	.81
Dependability	.83	.01	.82
Adjustment	.81	.02	.79
Cooperativeness	.68	.01	.67
Internal Control	.70	.02	.68
Physical Condition	.80	.01	.79

^aSpearman-Brown corrected.

^bR² is adjusted for shrinkage; N = 6310-6332.

^cSplit-half reliability minus ASVAB adjusted R².

The correlations among the ABLE composite scores are presented in Table 3.64. In general, the Achievement composite correlates the highest with the other composites; Physical Condition correlates the lowest with the other composites.

These composite scores also will not adversely impact women or minorities. Table 3.65 presents the means and standard deviations of the composites by gender and Table 3.66 presents the means and standard deviations by race. Gender differences are mixed and race differences tend to favor minorities.

In summary, the ABLE composite scores are reliable, they are independent of the ASVAB, and they will not adversely impact women or minorities.

Table 3.64**Longitudinal Validation: Correlations Among ABLE Composites**

Composite	Achievement	Leadership Potential	Dependability	Adjustment	Cooperativeness	Internal Control
Leadership Potential	.68					
Dependability	.62	.35				
Adjustment	.70	.54	.43			
Cooperativeness	.60	.40	.61	.54		
Internal Control	.56	.35	.54	.46	.46	
Physical Condition	.49	.36	.23	.37	.23	.19

Table 3.65**Longitudinal Validation: ABLE Composite Score Means and Effect Sizes by Gender**

Composite	Male		Female		Effect Size ^a (d)
	Mean	SD	Mean	SD	
Achievement Orientation	149.9	26.95	150.8	26.26	-.03
Leadership Potential	50.2	9.90	48.6	10.51	.16
Dependability	149.2	25.31	155.3	24.07	-.24
Adjustment	50.2	9.97	48.4	10.06	.18
Cooperativeness	49.8	10.00	51.0	9.98	-.12
Internal Control	49.7	10.14	52.1	8.77	-.24
Physical Condition	50.7	9.78	45.4	10.20	.54

Note. N for males = 5509-5529; N for females = 864-866.

^ad is the standardized mean difference between male and female scores. A positive value indicates superior performance by males; a negative value indicates superior performance by females.

Table 3.66

Longitudinal Validation: ABLE Composite Score Means and Effect Sizes by Race

Composite	White (N = 4424-4433)			Black (N = 1504-1514)			Hispanic (N = 221-223)			Other (N = 216-217)		
	Mean	SD	Effect Size ^a (d)	Mean	SD	Effect Size ^a (d)	Mean	SD	Effect Size ^a (d)	Mean	SD	Effect Size ^a (d)
Achievement Orientation	148.5	27.60	-.22	154.4	23.98	-.22	149.9	25.74	-.05	150.4	28.47	-.07
Leadership Potential	49.4	10.19	-.26	51.9	9.04	-.26	49.5	10.36	-.01	49.6	10.30	-.02
Dependability	148.4	26.10	-.23	154.1	21.96	-.23	155.3	23.26	-.26	149.8	26.51	-.05
Adjustment	49.5	10.31	-.21	51.5	9.01	-.21	50.2	9.41	-.07	49.9	9.86	-.04
Cooperativeness	49.3	10.08	-.27	52.1	9.54	-.27	50.3	9.70	-.10	49.2	9.73	.02
Internal Control	50.3	10.20	.08	49.5	9.26	.08	49.3	9.95	.09	48.5	10.48	.18
Physical Condition	49.5	10.20	-.23	51.7	9.25	-.23	49.4	9.57	.01	49.6	10.04	-.01

^ad is the standardized mean difference between two subgroup scores. All effect sizes in this table are relative to the white subgroup. A positive effect size indicates that whites score higher than the minority, and a negative value indicates that whites score lower.

SCORING AND FORMING COMPOSITES FOR THE AVOICE INVENTORY

This section begins with a brief review of the development and content of the AVOICE and then discusses the following topics: (a) data screening, (b) descriptive statistics based on the Initial Sample, (c) subgroup differences, (d) uniqueness estimates, and (e) composite formation.

Development and Content of the AVOICE Inventory

The AVOICE (Army Vocational Interest Career Examination) was developed to measure vocational interests relevant to jobs in the Army. The constructs measured by this inventory were selected through an extensive literature review reported in Hough (1986). The development and pilot testing of the AVOICE is described in Hough, Barge, and Kamp (1987); field testing is described in Hough, McGue, Houston, and Pulakos (1987); and the evaluation of the Revised Trial Battery version of the AVOICE is reported in Hough, McCloy, Ashworth, and Hough (in press).

Several modifications were made to the Trial Battery version of the AVOICE to form the Experimental Battery version. These changes were made in two phases, one phase resulting in the Revised Trial Battery version and the second in the Experimental Battery version.

To form the Revised Trial Battery version, numerous items were moved from one scale to another based on rational considerations, item-total scale correlations, factor analysis at the item level, clarity of interpretation, and practical considerations. In addition, numerous items were dropped because they also appeared on the ABLE, appeared on scales that were too long, or became "singletons" after other items were moved to different scales. Also, in two cases, two scales were merged to form a single scale: Teaching/Counseling was merged with Leadership to form Leadership/Guidance, and Armor/Cannon and Infantry were merged to form Combat.

Also, the Outdoor scale was dropped and several scales were renamed. Office Administration became Clerical/Administrative; Automated Data Processing was renamed Computers; Supply Administrative became Warehousing/Shipping; Marksman was renamed Firearms Enthusiast; Adventure was renamed Rugged Individualism; and Vehicle/Equipment Operator became Vehicle Operator. Finally, two scales were split into separate scales: the Law Enforcement scale was divided to form Law Enforcement and Fire Protection, and the Food Service scale was split to form Food Service Employee and Food Service Professional.

Up to this point, none of the changes entailed adding or modifying items. Thus, Revised Trial Battery scores can be obtained for the CV sample, which completed the Trial Battery version of the inventory.

In the second phase of changes, 16 items were added to increase the stability and reliability of the scales, one item was modified to better fit the AVOICE response format, and the instructions were modified slightly to allow for the use of separate answer sheets. The inventory that resulted is the Experimental Battery version that was administered in the Longitudinal Validation data collection.

Data Screening

First, cases were screened if more than 10 percent of their data was missing. Then, we investigated methods to detect careless or low-literacy respondents. Four indexes were selected: a chi-square index to detect patterned responding, a Runs index to detect repetitious responding, an Option Variance index to detect persons who tend to rely on very few of the (five) response options, and an empirically derived Unlikely Response scale. These indexes were developed by hypothesizing and measuring patterns of responses that would be produced only by careless or low-literacy respondents. The indexes are described in further detail in Figure 3.6.

Tables 3.67-3.70 report the frequency distributions for scores on these indexes in the Initial Longitudinal Sample, along with the cut score selected for flagging cases for deletion. The detection scales were newly developed, and we were concerned about erroneously removing inventories that had, in fact, been conscientiously completed. We, therefore, identified a conservative cut score for each index.

Having set these cut scores, we flagged respondents who scored beyond them. If flagged by one or more of the indexes, a case was removed from the sample. The results of this screen and the missing data screen are presented in Table 3.71, along with the screening results obtained in the CV sample (in which the ABLE Non-Random Response Scale and 10% missing data screens were applied). Overall, using the current method of data screening resulted in deleting a slightly smaller proportion of the sample. The chief advantage of the current screening procedure (over the use of the ABLE Non-Random Response scale) is that the screening indexes are based on responses to the AVOICE. (A person who responds carelessly to the ABLE will not necessarily respond carelessly to the AVOICE.)

For the inventories surviving these screens, missing data were treated in the following way. If more than 10 percent of the item responses in a scale were missing, the scale score was not computed; instead the scale score was treated as missing. If item responses were missing for a scale, but the percent missing was equal to or less than 10 percent, then the scale midpoint (3) was used in place of the missing response. The midpoint was chosen because the effect on the overall mean for the entire group would be less than if the average of the non-missing items in the scale were used.

Comparison of Descriptive Statistics for the Revised Trial Battery and Experimental Battery

Table 3.72 compares Revised Trial Battery (CV) and Experimental Battery (LV) AVOICE descriptive statistics, including means, standard deviations, median item-total correlations, and internal consistency (alpha) reliabilities. Test-retest reliabilities for the Revised Trial Battery are also presented (test-retest data were not collected in Longitudinal Validation).

Chi-Square Patterned Response Index

Definition:

A chi-square measure of independence of responses to items i and $i+1$ across all the items of the inventory.

Target Careless Responders:

Persons whose responses look like this: 1234512345 or 543543543543 or 1212121212, etc.

Runs Index

Definition:

The difference between the number of runs observed for the individual respondent and the average number in the total sample ($N = 6800$), where a run is a series of repeated item responses (e.g., 11111).

Target Careless Responders:

Persons whose responses look like this: 111111111111 or 222222222222 or 12325124134, etc. In the last example, the person never responded the same way twice in a row, which would be extremely rare for a careful responder but perhaps more common among careless responders.

Option Variance Index

Definition:

The number of times an individual selected each of the five options was determined. The index is the variance of these five frequencies. High variance reflects a tendency to use a small subset of response options.

Target Careless Responders:

Persons who consistently select one or two of the response options. Their response patterns might look like this: 12221112222112212221 or 3331111133113331113 or 22222221122222211322, etc. Careful responders should not produce patterns of this nature.

Unlikely Response Scale

Definition:

The number of times an individual selects a response option that was selected by fewer than 5% of the sample. The response options were selected such that they are evenly distributed across the scales of each inventory.

Target Careless Responders:

Anyone who fails to read and respond carefully to the questions.

Figure 3.6. Longitudinal Validation: Screening indexes developed for the AVOICE.

Table 3.67

Frequency Distribution of Scores on the AVOICE Chi-Square Patterned Response Index (With Cut Score)

Score Interval Midpoint ^a		Frequency ^b	Percent	Cumulative Percent
0	*****	812	11.8	11.8
25	*****	4330	63.1	75.0
50	*****	1123	16.4	91.3
75	**	317	4.6	96.0
100	*	129	1.9	97.8
125	*	70	1.0	98.9
150	* ----- CUT SCORE > 137.5 -----	26	.4	99.2
⋮		⋮		⋮
475		1	.0	100.0

10 20 30 40 50 60
 Percent

^aNot a true midpoint for the first and last intervals.

^bThe Missing Data screen was applied prior to this analysis. Thus, persons who responded to fewer than 90% of the items were not included in this analysis.

Table 3.68

Frequency Distribution of Scores on the AVOICE Runs Index (With Cut Score)

Score Interval Midpoint ^a		Frequency ^b	Percent	Cumulative Percent
0	*****	815	11.9	11.9
6	*****	1603	23.4	35.3
12	*****	1421	20.7	56.0
18	*****	1112	16.2	72.2
24	*****	841	12.3	84.4
30	*****	402	5.9	90.3
36	****	191	2.8	93.1
42	**	107	1.6	94.7
48	**	92	1.3	96.0
54	*	76	1.1	97.1
60	* ----- CUT SCORE > 57 -----	43	.6	97.7
⋮		⋮		⋮
108		9	.1	100.0

3 6 9 12 15 18 21
 Percent

^aNot a true midpoint for the first and last intervals.

^bThe Missing Data screen was applied prior to this analysis. Thus, persons who responded to fewer than 90% of the items were not included in this analysis.

Table 3.69

Frequency Distribution of Scores on the AVOICE Option Variance Index (With Cut Score)

Score Interval Midpoint ^a		Frequency ^b	Percent	Cumulative Percent
0.016	*****	3551	51.8	51.8
0.032	*****	1376	20.1	71.8
0.048	*****	737	10.7	82.6
0.064	***	418	6.1	88.7
0.080	**	207	3.0	91.7
0.096	*	142	2.1	93.8
0.112	*	109	1.6	95.3
0.128	*	82	1.2	96.5
0.144	*----- CUT SCORE > .136 -----	56	.8	97.4
⋮		⋮		⋮
0.320		3	.0	100.0

10 20 30 40 50
Percent

^aNot a true midpoint for the first and last intervals.

^bThe Missing Data screen was applied prior to this analysis. Thus, persons who responded to fewer than 90% of the items were not included in this analysis.

Table 3.70

Frequency Distribution of Scores on the AVOICE Unlikely Response Scale (With Cut Score)

Score		Frequency ^a	Percent	Cumulative Percent
0	*****	3531	51.5	51.5
1	*****	1621	23.6	75.1
2	*****	725	10.6	85.7
3	***	417	6.1	91.8
4	**	217	3.2	94.9
5	*	132	1.9	96.9
6	*----- CUT SCORE > 5 -----	74	1.1	97.9
⋮		⋮		⋮
12		12	.2	100.0

10 20 30 40 50
Percent

^aThe Missing Data screen was applied prior to this analysis. Thus, persons who responded to fewer than 90% of the items were not included in this analysis.

Table 3.71**Comparison of CV and LV AVOICE Data Screening Results**

	<u>Number</u>		<u>Percent</u>	
	CV	LV	CV	LV
Number of Inventories Scanned	9359	7000	100.0	100.0
Deleted Using Overall Missing Data Screen (Decision Rule: If missing data greater than 10%, delete inventory)	200	141	2.1	2.0
Deleted by at least one of the four response validity screens (LV Battery only) or by the ABLE Non-Random Response Scale (CV only)	760	527	8.1	7.5
Deleted Using Chi-Square Patterned Response Index Screen	---	78	---	1.1
Deleted Using Runs Index Screen	---	199	---	2.8
Deleted Using Option Variance Index Screen	---	237	---	3.4
Deleted Using Unlikely Response Scale Screen	---	216	---	3.1
Respondents Passing Screening Criteria	8399	6332	89.7	90.5

Several findings are noteworthy. In general, the LV sample tended to score higher on most of the scales, especially Combat, Law Enforcement, Firearms Enthusiast, Food Service - Employee, and Fire Protection. Where mean scores increased, standard deviations tended to decline. Still, internal consistency reliabilities all remained quite high, ranging from .78 to .95. Adding items to some scales produced the expected increase in reliability.

Analysis of Subgroup Differences

Means and standard deviations for the AVOICE scales by gender are shown in Table 3.73. Mean scores for men exceeded the means for women on 13 of the 22 scales. In particular, men tended to score higher on Mechanics, Heavy Construction, Electronics, Combat, Rugged Individualism, and Firearms Enthusiast. Women scored higher on Clerical/Administrative, Medical Services, and Aesthetics.

Table 3.72

Comparison of AVOICE Scales Scores and Reliabilities for the Revised Trial (CV) and Experimental (LV) Batteries

AVOICE Scale	No. Items		Sample Size		Mean		SD		Effect Size ^b		Median Item-Total Correlation ^c		Internal Consistency Reliability (Alpha)		Test-Retest Reliability ^d	
	CV	LV	CV	LV	CV	LV	CV	LV	CV	LV	CV	LV	CV	LV	CV	LV
Clerical/Admin	14	14	8463	6252	39.6	40.0	10.81	10.23	.04		.67	.64	.92	.92	.78	.78
Mechanics	10	10	8382	6315	32.1	32.9	9.42	9.38	.09		.80	.79	.94	.95	.82	.82
Heavy Construction	13	13	8488	6302	39.3	38.7	10.54	9.66	-.06		.68	.63	.92	.91	.84	.84
Electronics	12	12	8359	6313	38.4	37.8	10.22	9.65	-.06		.70	.68	.94	.93	.81	.81
Combat	10	10	8466	6318	26.5	33.8	8.35	7.48	.98		.65	.61	.90	.88	.73	.73
Medical Services	12	12	8364	6263	36.9	37.4	9.54	9.42	.05		.68	.68	.92	.91	.78	.78
Rugged Individualism	15	16	8396	6316	53.3	59.2	11.44	10.38	.23		.58	.52	.90	.88	.81	.81
Leadership/Guidance	12	12	8444	6259	40.1	41.7	8.63	8.32	.19		.62	.61	.89	.89	.72	.72
Law Enforcement	8	8	8471	6317	24.7	26.8	7.37	6.73	.31		.65	.62	.89	.87	.84	.84
Food Service - Prof	8	8	8472	6308	20.2	20.9	6.50	6.39	.11		.67	.63	.89	.87	.75	.75
Firearms Enthusiast	7	7	8397	6316	23.0	25.1	6.36	5.78	.36		.66	.63	.89	.88	.80	.80
Science/Chemical	6	6	8468	6318	16.9	17.1	5.33	4.96	.04		.70	.60	.85	.82	.74	.74
Drafting	6	6	8493	6320	19.4	19.4	4.97	4.94	.00		.66	.65	.84	.83	.74	.74
Autographics	5	5	8473	6251	17.6	17.3	4.09	3.83	-.08		.69	.62	.83	.79	.75	.75
Aesthetics	5	5	8413	6228	14.2	14.4	4.13	4.12	.05		.59	.56	.79	.78	.73	.73
Computers	4	4	8224	6266	14.0	13.1	3.99	4.03	-.22		.78	.77	.90	.89	.77	.77
Food Service - Empl	3	6	8304	6318	5.1	12.2	2.08	4.39	.46		.54	.66	.73	.85	.56	.56
Mathematics	3	3	8421	6281	9.6	9.3	3.09	3.00	-.10		.78	.73	.88	.85	.75	.75
Electronic Comm	6	6	8403	6272	18.4	19.9	4.66	4.21	.36		.60	.58	.83	.81	.68	.68
Warehousing/Shipping	2	7	8407	6321	5.8	20.4	1.75	5.02	.02		.44	.62	.61	.85	.54	.54
Fire Protection	2	6	8431	6320	6.1	19.8	1.96	4.37	.34		.62	.55	.76	.81	.67	.67
Vehicle Operator	3	6	8378	6320	8.8	17.8	2.65	4.54	.04		.51	.55	.70	.78	.68	.68

^aInitial longitudinal sample screened for missing data and random responding.^bEffect Size = $\frac{\text{Mean}_{LV} - ([\# \text{ items}_{LV} / \# \text{ items}_{CV}] \times \text{Mean}_{CV})}{\text{Standard Deviation}_{LV}}$. Positive effect sizes indicate higher mean scores in the LV sample.^cEach item was correlated with the sum of the remaining items in the scale.^d $d_M = 389-409$ for test-retest correlations.

Table 3.73**Longitudinal Validation: AVOICE Scale Score Means and Effect Sizes by Gender**

AVOICE Scale	Male (N = 5450-5530)		Female (N = 788-802)		Effect Size ^a (d)
	Mean	SD	Mean	SD	
Clerical/Administrative	39.0	9.73	46.6	11.20	-.77
Mechanics	33.7	9.01	27.4	10.06	.69
Heavy Construction	39.7	9.21	31.4	9.65	.90
Electronics	38.7	9.34	31.9	9.83	.72
Combat	34.7	7.04	27.9	7.85	.95
Medical Services	36.9	9.14	41.3	10.44	-.48
Rugged Individualism	60.2	9.90	51.7	10.76	.85
Leadership/Guidance	41.5	8.27	43.1	8.61	-.20
Law Enforcement	27.0	6.64	25.6	7.26	.21
Food Service - Professional	20.7	6.31	22.2	6.82	-.24
Firearms Enthusiast	25.9	5.32	9.6	5.91	1.17
Science/Chemical	17.3	4.90	16.1	5.27	.24
Drafting	19.6	4.89	18.2	5.13	.28
Audiographics	17.2	3.84	18.0	3.72	-.19
Aesthetics	14.1	4.03	16.8	3.98	-.67
Computers	13.1	4.02	13.3	4.14	-.07
Food Service - Employee	12.2	4.32	12.4	4.82	-.06
Mathematics	9.3	2.96	9.7	3.26	-.16
Electronic Communications	19.9	4.15	19.8	4.63	.01
Warehousing/Shipping	20.5	4.94	.3	5.55	.03
Fire Protection	20.0	4.25	18.2	4.88	.42
Vehicle Operator	18.0	4.48	16.4	4.76	.36

^ad is the standardized mean difference between male and female scores. A positive value indicates superior performance by males; a negative value indicates superior performance by females.

Means and standard deviations for these scales by race are shown in Table 3.74. In general, minorities tended to score higher than whites on these scales. Mean scores for blacks, for instance, were higher than those for whites on 15 of the 22 scales. On the average, Hispanics scored higher than whites on 15 of the 22 scales, and other minorities scored higher than whites on 12 of the scales.

Uniqueness Analysis

The Uniqueness estimates for the AVOICE scales are shown in Table 3.75. Compared to the ABLE, the AVOICE shares more variance with the ASVAB. Clerical/Administrative, Mechanics, Heavy Construction, Electronics, Firearms Enthusiast, and Mathematics overlap the most with the ASVAB. Nevertheless, the AVOICE scales still measure a high amount of unique and reliable variance.

Formation of AVOICE Composites

Like the ABLE, the AVOICE is composed of scales intended to measure unique constructs; overlap between scales was avoided so that each scale might make a unique contribution in the prediction of job performance. Given the large number of predictors available in the Longitudinal Validation, however, we attempted to identify clusters of AVOICE scales that could be combined to form composite scores, thus reducing the number of predictors.

Correlations among the AVOICE scales appear in Table 3.76. As expected, most of these correlations are quite low, indicating that we were relatively successful in measuring independent areas of vocational interest. Still, several clusters of related scales do seem to appear in this matrix. For example, Firearms Enthusiast correlates .69 with Combat and .72 with Rugged Individualism. Other correlated pairs of scales are more difficult to interpret, however.

Principal components analysis was conducted to identify sets of AVOICE scales that cluster together empirically. The method of parallel analysis (Humphreys & Montanelli, 1975; Montanelli & Humphreys, 1976) indicated that as many as 22 components underlie the 22 scales, highlighting the difficulty of clustering scales intended to measure different constructs. Through a series of factor analyses, however, several clusters or pairs of scales appeared consistently. Ten different models for combining the scales were hypothesized based on these analyses, and these models were compared using LISREL (Joreskog & Sorbom, 1986). Table 3.77 presents four of these models and the results obtained for each model. The model on the far right was selected for its superior fit to the data and the interpretability of the model's composites.

The CV and LV composite formation models for AVOICE are presented side-by-side in Figure 3.7. As shown, there are eight LV composites and six CV composites; the CV Skilled Technical composite has been separated into three more homogeneous composites--Interpersonal, Administrative, and Skilled/Technical.

Table 3.74

Longitudinal Validation: AVOICE Scale Score Means and Effect Sizes by Race

AVOICE Scale	White (N = 4447-4404)			Black (N = 1455-1487)			Hispanic (N = 215-219)			Other (N = 216-217)		
	Mean	SD	Effect Size ^a (d)	Mean	SD	Effect Size ^a (d)	Mean	SD	Effect Size ^a (d)	Mean	SD	Effect Size ^a (d)
Clerical/Administrative	37.6	9.59		46.6	9.20	-.95	41.8	9.36	-.44	41.2	10.64	-.37
Mechanics	33.1	9.61	.10	32.2	8.88	.10	33.1	7.91	.00	33.8	9.45	-.07
Heavy Construction	39.0	9.67	.09	38.1	9.62	.09	36.7	9.37	.24	37.4	9.77	.17
Electronics	36.8	9.49	-.37	40.3	9.68	-.37	40.0	9.18	-.34	39.7	9.84	-.31
Combat	34.5	7.38	.35	31.9	7.53	.35	33.6	6.96	.12	33.5	7.81	.14
Medical Services	36.4	9.41	-.42	40.3	8.82	-.42	39.1	9.41	-.29	38.1	9.53	-.18
Rugged Individualism	61.6	9.42	.98	52.2	10.24	.98	57.8	9.37	.40	58.2	9.70	.36
Leadership/Guidance	40.9	8.34	-.34	43.7	7.83	-.34	42.1	8.37	-.14	42.2	9.02	-.16
Law Enforcement	27.3	6.78	.25	25.6	6.48	.25	26.0	6.38	.14	25.9	6.55	.21
Food Service - Professional	20.2	6.25	-.48	23.2	6.41	-.48	20.3	5.80	-.02	20.2	6.20	.00
Firearms Enthusiast	25.8	5.77	.47	23.1	5.51	.47	24.0	5.06	.31	24.4	5.47	.24
Science/Chemical	17.0	4.87	.00	17.3	5.10	.00	18.4	5.02	-.29	17.9	5.40	-.18
Drafting	19.3	4.96	-.04	19.5	4.84	-.04	20.1	4.95	-.16	20.9	5.18	-.32
Audio/graphics	17.0	3.86	-.37	18.4	3.59	-.37	17.9	3.54	-.23	17.6	3.93	-.16
Aesthetics	13.9	4.10	-.39	15.5	3.93	-.39	15.5	3.74	-.39	15.3	4.34	-.34
Computers	12.3	4.02	-.73	15.1	3.31	-.73	14.6	3.66	-.57	14.1	3.81	-.45
Food Service - Employee	11.7	4.12	-.49	13.8	4.83	-.49	12.4	4.03	-.17	11.7	4.26	.00
Mathematics	9.0	2.98	-.41	10.2	2.87	-.41	10.2	2.93	-.40	10.1	2.97	-.37
Electronic Communications	19.4	4.13	-.44	21.2	4.15	-.44	21.0	4.14	-.39	20.4	4.36	-.24
Warehousing/Shipping	19.8	4.89	-.53	22.4	4.91	-.53	20.8	4.77	-.20	19.8	5.11	.00
Fire Protection	19.9	4.32	.14	19.3	4.53	.14	20.0	4.36	-.02	19.8	4.15	.02
Vehicle Operator	17.7	4.51	-.15	18.4	4.58	-.15	17.3	4.46	.09	16.7	4.64	.22

^ad is the standardized mean difference between two subgroup scores. All effect sizes in this table are relative to the white subgroup. A positive effect size indicates that whites score higher than the minority, and a negative value indicates that whites score lower.

Table 3.75

Comparison of Reliability Coefficients, Multiple Regression Coefficients, and Uniqueness Estimates for AVOICE Scale Scores: CV/LV

AVOICE Scale	No. Items	Alpha Coefficient ^a	ASVAB R-Squared ^b	Uniqueness ^c
Clerical Administrative	14/14	.92/.92	.13/.14	.79/.78
Mechanics	10/10	.94/.95	.20/.18	.74/.77
Heavy Construction	13/13	.92/.91	.16/.13	.76/.78
Electronics	12/12	.94/.93	.10/.08	.84/.85
Combat	10/10	.90/.88	.02/.06	.88/.82
Medical Services	12/12	.92/.91	.05/.03	.87/.88
Rugged Individualism	15/16	.90/.88	.16/.20	.74/.68
Leadership/Guidance	12/12	.89/.89	.04/.03	.85/.86
Law Enforcement	8/8	.89/.87	.01/.01	.88/.86
Food Service-Professional	8/8	.89/.87	.04/.03	.85/.84
Firearms Enthusiast	7/7	.89/.88	.14/.16	.75/.72
Science/Chemical	6/6	.85/.82	.03/.03	.82/.79
Drafting	6/6	.84/.83	.02/.03	.82/.80
Audiographics	5/5	.83/.79	.02/.01	.81/.78
Aesthetics	5/5	.79/.78	.07/.06	.72/.72
Computers	4/4	.90/.89	.04/.06	.86/.83
Food Service-Employee	3/6	.73/.85	.03/.04	.70/.81
Mathematics	3/3	.88/.85	.15/.16	.73/.69
Electronic Communication	6/6	.83/.81	.01/.01	.82/.80
Warehousing/Shipping	2/7	.61/.85	.05/.04	.56/.81
Fire Protection	2/6	.76/.81	.02/.03	.74/.78
Vehicle Operator	3/6	.70/.78	.09/.06	.61/.72

^aN = 8325/6251.

^bR² is adjusted for shrinkage (i.e., cross-validity estimated); N = 6381/4930.

^cAlpha reliability minus ASVAB adjusted R².

Table 3.76

Longitudinal Validation: Correlations Among 22 AVOICE Scale Scores

AVOICE Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. Clerical/Administrative																					
2. Mechanics	.03																				
3. Heavy Construction	.06	.60																			
4. Electronics	.29	.62	.53																		
5. Combat	-.03	.40	.49	.35																	
6. Medical Services	.48	.00	.06	.21	.06																
7. Rugged Individualism	-.19	.37	.48	.29	.64	.07															
8. Leadership/Guidance	.44	.04	.11	.27	.26	.51	.24														
9. Law Enforcement	.09	.14	.24	.10	.40	.19	.32	.28													
10. Food Serv.-Professional	.50	.09	.18	.16	.03	.30	-.03	.27	.08												
11. Firearms Enthusiast	-.08	.44	.49	.36	.69	.04	.72	.19	.38	.01											
12. Science/Chemical	.33	.21	.27	.47	.35	.44	.25	.37	.18	.18	.24										
13. Drafting	.27	.21	.29	.46	.26	.25	.29	.34	.06	.17	.23	.39									
14. Audiographics	.44	.14	.15	.40	.13	.40	.09	.39	.09	.32	.11	.35	.43								
15. Aesthetics	.43	-.03	.00	.17	.02	.43	.06	.46	.02	.32	-.02	.31	.30	.37							
16. Computers	.54	.16	.10	.53	.07	.31	-.07	.31	.01	.18	.02	.40	.34	.37	.25						
17. Food Service - Employee	.45	.13	.25	.15	.07	.25	-.05	.14	.11	.73	.02	.19	.09	.19	.25	.15					
18. Mathematics	.52	.11	.13	.35	.08	.31	.02	.41	.06	.24	.03	.33	.31	.23	.27	.45	.21				
19. Electronic Communication	.47	.31	.28	.63	.37	.37	.24	.46	.16	.22	.27	.48	.45	.50	.33	.51	.18	.39			
20. Warehousing/Shipping	.64	.26	.40	.36	.21	.29	.07	.33	.16	.49	.18	.29	.24	.37	.24	.33	.48	.35	.41		
21. Fire Protection	.16	.26	.44	.33	.45	.35	.45	.36	.48	.16	.42	.35	.25	.28	.16	.12	.14	.13	.35	.32	
22. Vehicle Operator	.27	.42	.53	.31	.31	.11	.21	.14	.28	.36	.29	.14	.10	.28	.05	.10	.39	.14	.24	.54	.35

Note. N = 6170-6332.

Table 3.77

Longitudinal Validation: Comparison of Four AVOICE Composite Formation Models, Using LISREL

	Model 1	Model 2	Model 3	Model 4
Degrees of Freedom:	199	194	199	183
Chi-Square:	8065	7323	7424	6716
Goodness-of-Fit Index:	.845	.859	.857	.871
Adjusted Goodness-of-Fit:	.803	.816	.818	.821
Root Mean Square Residual:	.239	.279	.244	.225
Composite Formation Model:	Electronics Science/Chemical Computers Mathematics Electronic Communication Clerical/Administrative Mechanics Heavy Construction Combat Rugged Individualism Firearms Enthusiast Food Service - Profession Food Service - Employee Warehousing/Shipping Vehicle Operator Law Enforcement Fire Protection Medical Services Leadership/Guidance Drafting Audiographics Aesthetics	Clerical/Administrative Medical Services Leadership/Guidance Science/Chemical Computers Mathematics Electronic Communications Electronics Combat Rugged Individualism Firearms Enthusiast Mechanics Heavy Construction Vehicle Operator Food Service - Employee Food Service - Professional Warehousing/Shipping Law Enforcement Fire Protection Drafting Audiographics Aesthetics	Combat Rugged Individualism Law Enforcement Fire Protection Firearms Enthusiast Medical Services Leadership/Guidance Science/Chemical Drafting Audiographics Aesthetics Electronic Communication Food Service - Professional Food Service - Employee Warehousing/Shipping Mechanics Heavy Construction Electronics Vehicles Operator Clerical/Administrative Computer Mathematics	Combat Rugged Individualism Firearms Enthusiast Drafting Audiographics Aesthetics Medical Services Leadership/Guidance Science/Chemical Computers Mathematics Electronic Communications Clerical/Administrative Warehousing/Shipping Food Service - Employee Food Service - Professional Law Enforcement Fire Protection Mechanics Heavy Construction Electronics Vehicle Operator

Longitudinal Validation Composites**Rugged/Outdoors**

Combat
Rugged Individualism
Firearms Enthusiast

Audiovisual Arts

Drafting
Audiographics
Aesthetics

Interpersonal

Medical Services
Leadership/Guidance

Skilled/Technical

Science/Chemical
Computers
Mathematics
Electronic Communications

Administrative

Clerical/Administrative
Warehousing/Shipping

Food Service

Food Service - Professional
Food Service - Employee

Protective Services

Fire Protection
Law Enforcement

Structural/Machines

Mechanics
Heavy Construction
Electronics
Vehicle Operator

Concurrent Validation Composites**Combat-Related**

Combat
Rugged Individualism
Firearms Enthusiast

Audiovisual Arts

Drafting
Audiographics
Aesthetics

Skilled/Technical

Clerical/Administrative
Medical Services
Leadership/Guidance
Science/Chemical
Data Processing
Mathematics
Electronic Communications

Food Service

Food Service - Professional
Food Service - Employee

Protective Services

Law Enforcement
Fire Protection

Structural/Machines

Mechanics
Heavy Construction
Electronics
Vehicle/Equipment Operator

Note. Warehousing/Shipping was not included in a CV composite.

Figure 3.7 Comparison of AVOICE composites for the Longitudinal and Concurrent Validations.

The split-half reliabilities for the LV composites, along with the ASVAB uniquenesses, are shown in Table 3.78. These reliabilities were estimated by creating split-half scale scores, then forming and correlating composite halves. All reliabilities and uniquenesses are quite high. Correlations among the AVOICE composites are shown in Table 3.79. Most of these correlations are quite low. Thus the AVOICE composites each reliably measure a relatively unique domain of vocational interest and each composite has potential to contribute to the prediction of job performance when used in conjunction with the ASVAB.

Descriptive statistics by gender and race are shown in Tables 3.80 and 3.81, respectively. Gender differences are mixed: Men tend to score higher on Rugged/Outdoors, Protective Services, and Structural/Machines, whereas women score higher on Audiovisual Arts, Interpersonal, and Administrative. Minority groups tend to score higher than whites on most of these composite scores, although whites tend to score higher than minority groups on Rugged/Outdoors and Protective Services.

SCORING AND FORMING COMPOSITES FOR THE JOB INVENTORY

After a brief review of the development and content of the JOB, this section covers the following: (a) data screening, (b) descriptive statistics for the Initial Sample, (c) subgroup differences, (d) uniqueness estimates, and (e) composite formation.

Development and Content of the JOB Inventory

The JOB (Job Orientation Blank) was developed to measure work environment preferences. Like the ABLE and the AVOICE, this inventory has been developed iteratively and has undergone changes over a period of several years. The constructs measured were selected as the result of an extensive literature review, which is reported in Hough (1986). The evaluation of the Trial Battery and Revised Trial Battery versions of the JOB is reported in Hough, McCloy, Ashworth, and Hough (1987).

Several changes have been made to the Trial Battery version of the JOB to form the Experimental Battery version. These changes were made in two phases. The first phase resulted in a revised method for scoring the JOB, which we used to score the CV data. We use the term Revised Trial Battery to refer to CV data scored with this method. The second phase of changes was based on data analyses of the pretest of the Experimental Battery. The changes for both phases are described below.

Phase One. We revised the method of scoring the JOB for the CV data on the basis of item-total scale correlations, factor analyses at the item level, clarity of interpretation, and practical considerations.

Factor analyses (principal factor with varimax rotation) of the JOB scales resulted in two factors that had little similarity with the six original constructs. We therefore investigated the structure of the JOB at the item level to learn more about what the JOB was measuring. We factor analyzed the 38 JOB items (principal factor with varimax rotation) and

Table 3.78**Longitudinal Validation: AVOICE Composite Score Reliability and Uniqueness Estimates**

Composite	Split-Half Reliability ^a	ASVAB R-Squared ^b	Uniqueness ³
Rugged/Outdoors	.91	.16	.75
Audiovisual Arts	.85	.02	.83
Interpersonal	.93	.04	.89
Skilled/Technical	.92	.05	.87
Administrative	.93	.09	.84
Food Service	.91	.04	.87
Protective Services	.85	.02	.83
Structural/Machines	.93	.14	.79

^aSpearman-Brown corrected.^bR² is adjusted for shrinkage; N = 6104-6261.^cSplit-half reliability minus ASVAB adjusted R².**Table 3.79****Longitudinal Validation: Correlations Among AVOICE Composites**

Composite	Rugged/ Outdoors	Audiovisual Arts	Inter- personal	Skilled/ Technical	Adminis- trative	Food Service	Protective Services
Audiovisual Arts	.19						
Interpersonal	.18	.57					
Skilled/Technical	.23	.59	.57				
Administrative	.03	.48	.49	.59			
Food Service	.01	.32	.30	.28	.57		
Protective Services	.53	.22	.40	.26	.24	.15	
Structural/Machines	.53	.30	.17	.43	.38	.29	.39

Note: N = 6080-6319.

Table 3.80

Longitudinal Validation: AVOICE Composite Score Means and Effect Sizes by Gender

Composite	Male		Female		Effect Size ^a (d)
	Mean	SD	Mean	SD	
Rugged/Outdoors	153.6	24.64	125.5	27.00	1.13
Audiovisual Arts	149.3	22.77	155.2	22.84	-.26
Interpersonal	99.2	17.14	105.9	17.92	-.39
Skilled/Technical	200.0	30.13	200.1	30.76	.00
Administrative	99.1	17.69	106.3	19.52	-.40
Food Service	99.6	18.31	102.6	20.15	-.16
Protective Services	100.8	16.77	94.6	18.97	.36
Structural/Machines	203.3	29.93	177.4	34.16	.85

Note. N for males = 5385-5525; N for females = 784-802.

^ad is the standardized mean difference between male and female scores.

A positive value indicates higher scores for males; a negative value indicates higher scores for females.

obtained three factors. The first factor consisted of positive work environment characteristics; the second factor consisted of negative work environment characteristics; the third factor consisted of items describing preferences for autonomous work settings or environments. The JOB was intended to measure preferences for work environments that allow for achievement, safety, comfort, status, altruism, and/or autonomy. The JOB was not measuring the intended work environment constructs; only the autonomy scale appeared reasonable.

Considerable prior research by Dawis and Lofquist (1984) indicates that the constructs have merit and are measurable. We concluded that the present items were not good measures of the constructs. We speculated that perhaps the reading level of the negatively worded items was too high for the present sample. If this were true, factor analyzing only the simply stated (positively stated) items might result in a more meaningful structure. We investigated this possibility by factor analyzing 29 of the 38 items. Six meaningful factors merged: Job Pride, Job Security, Serving Others, Job Autonomy, Job Routine, and Ambition. We reconstituted the JOB scales according to these six factors. Revised Trial Battery JOB scale data are reported for these six factors.

We incorporated these changes into the pretest version of the JOB Experimental Battery scales. We also revised five negatively worded items that had been deleted in the factor analysis mentioned above and included them in the pretest version of the JOB Experimental Battery inventory.

Table 3.81

Longitudinal Validation: AVOICE Composite Score Means and Effect Sizes by Race

Composite	White (N = 1504-1514)			Black (N = 221-223)			Hispanic (N = 4424-4433)			Other (N = 216-217)		
	Mean	SD		Mean	SD		Mean	SD		Mean	SD	Effect Size ^a (d)
Rugged/Outdoors	154.5	25.55		137.4	26.22	.67	146.6	23.70	.31	147.6	25.77	.27
Audiovisual Arts	147.6	22.78		155.5	21.83	-.35	155.5	21.69	-.35	155.7	24.68	-.35
Interpersonal	98.0	17.20		105.6	16.49	-.45	102.3	17.62	-.25	101.2	18.37	-.19
Skilled/Technical	195.2	29.61		211.2	28.13	-.55	211.8	29.92	-.56	207.8	31.33	-.43
Administrative	96.4	17.23		110.4	16.58	-.82	102.6	16.86	-.36	99.8	18.71	-.20
Food Service	97.7	17.72		107.2	19.57	-.52	99.6	16.77	-.11	97.8	18.02	-.01
Protective Services	101.1	17.07		97.1	17.23	.23	99.3	17.45	.10	98.6	16.68	.14
Structural/Machines	199.2	31.73		202.6	31.75	-.11	199.3	29.19	.00	199.4	32.22	-.01

^ad is the standardized mean difference between two subgroups' scores. All effect sizes in this table are relative to the white subgroup. A positive effect size indicates that whites score higher than the minority, and a negative value indicates that whites score lower.

Phase Two. The JOB Experimental Battery pretest version was administered to 57 AIT students (Army enlisted soldiers) at Fort Belvoir in May 1986. We examined the item-total scale correlations and dropped five items from the JOB as a result of low item-total scale correlations. We computed alpha coefficients and obtained internal consistency reliabilities of .74, .60, .56, .47, and .68 for Job Pride, Job Security/Comfort, Serving Others, Job Autonomy, Job Routine, and Ambition, respectively. The reliability for the JOB scale Ambition was considerably better than the prior version.

We incorporated the above-described changes into the JOB Experimental Battery and added two items, one to the Job Routine scale and one to the Serving Others scale to increase the number of items in each scale to improve the reliabilities of those scales. We also changed the response option "Indifferent" to "Doesn't Matter". The inventory that resulted is the JOB Experimental Battery which consists of 31 items and the following six scales: Job Pride, Job Security/Comfort, Serving Others, Job Autonomy, Job Routine, and Ambition. The Experimental Battery version of the JOB was administered to the LV sample.

Data Screening

JOB scores were deleted when more than 10 percent of the item responses were missing from a person's data. Then we investigated methods for screening careless or low-literacy respondents. The JOB, like the AVOICE, contains no scales designed to detect such persons, so we developed screening indexes similar to those developed for the AVOICE (see Figure 3.5). They were developed by hypothesizing and quantitatively capturing patterns of responses that might be produced only by persons who either respond carelessly or do not understand the questions. A cut score was established for each index at the extreme of the distribution, resulting in relatively few persons being screened by each index.

Tables 3.82-3.85 display frequency distributions for scores on these indexes in the Initial Sample, along with cut scores selected for flagging cases for deletion. Cases were removed from the sample if flagged by any one of the indexes. Table 3.86 reports the number and proportion of persons screened from the LV and CV samples. In the CV sample, two screens were used: the 10 percent missing data rule and the ABLE Non-Random Response scale screen. Fewer persons were screened by the missing data screen in the LV sample. The screening indexes developed for the JOB removed about the same proportion of respondents from the LV sample. Overall, a smaller proportion was deleted from the LV sample, compared to the CV sample.

For the inventories surviving these screens, missing data were treated in the following way. If more than 10 percent of the item responses in a scale were missing, the scale score was not computed; instead the scale score was treated as missing. If item responses were missing for a scale but the percent missing was equal to or less than 10 percent, then the midpoint of the scale (3) was used in place of the missing response.

Longitudinal Validation: Frequency Distribution of Scores on the JOB Chi-Square Patterned Response Index (With Cut Score)

^bThe Missing Data screen was applied prior to this analysis. Thus, persons who responded to fewer than 90% of the items were not included in this analysis.

Longitudinal Validation: Frequency Distribution of Scores on the JOB Runs Test (With Cut Score)

- *Not a true midpoint for the first and last intervals.
- *The Missing Data screen was applied prior to this analysis. Thus, persons who responded to fewer than 90% of the items were not included in this analysis.

Longitudinal Validation: Frequency Distribution of Scores on the JOB Option Variance Index (With Cut Score)

^bThe Missing Data screen was applied prior to this analysis. Thus, persons who responded to fewer than 90% of the items were not included in this analysis.

Longitudinal Validation: Frequency Distribution of Scores on the JOB Unlikely Response Scale (With Cut Score)

*The Missing Data screen was applied prior to this analysis. Thus, persons who responded to fewer than 90% of the items were not included in this analysis.

Table 3.86**Comparison of CV and LV JOB Data Screening Results**

	Number		Percent	
	CV	LV	CV	LV
Number of Inventories Scanned	9,359	7,000	100.0	100.0
Deleted Using Overall Missing Data Screen (Decision Rule: If missing data greater than 10%, delete inventory)	378	119	4.0	1.7
Deleted by at least one of the four response validity screens (Longitudinal Validation only) or by the ABLE Non-Random Response Scale (Concurrent Validation)	742	559	7.9	8.0
Deleted Using Chi-Square Patterned Response Index Screen		117		1.7
Deleted Using Runs Index Screen		251		3.6
Deleted Using Option Variance Index Screen		145		2.1
Deleted Using Unlikely Response Scale Screen		113		1.6
Respondents Passing Screening Criteria	8,239	6,322	88.1	90.3

**Comparison of Descriptive Statistics for the
Revised Trial Battery and Experimental Battery**

Table 3.87 compares Revised Trial Battery (CV) and Experimental Battery (LV) JOB scale score descriptive statistics, including means, standard deviations, and reliabilities. LV respondents scored substantially higher than CV respondents on two of the scales -- Job Security/Comfort and Job Routine -- and lower on Job Autonomy. In general, internal consistency (alpha) reliabilities are higher in the LV sample, compared to the CV sample, with estimates ranging from .59 to .80. These are fairly high reliabilities for scales as short as these.

Analysis of Subgroup Differences

The means and standard deviations for the JOB scales by gender are presented in Table 3.88. On the average, women score higher than men on four of the six scales. In particular, they tend to value serving others and job security more than men do.

Table 3.87

Comparison of JOB Scale Scores and Reliabilities for Revised Trial (CV) and Experimental (LV) Batteries

JOB Scale	No. Items		Sample Size		Mean		SD		Effect Size ^a	Median Item-Total Correlation ^b		Internal Consistency Reliability (Alpha)	
	CV	LV	CV	LV	CV	LV	CV	LV		CV	LV	CV	LV
Job Pride	10	10	7809	6309	43.6	44.1	4.51	4.01	.12	.54	.51	.84	.79
Job Security/Comfort	5	6	7817	6322	21.6	27.1	2.33	2.41	.49	.43	.52	.67	.76
Serving Others	3	3	7784	6290	12.1	12.1	1.83	2.02	.00	.52	.63	.66	.80
Job Autonomy	4	4	7817	6228	15.1	14.5	2.29	2.41	-.25	.31	.36	.50	.59
Job Routine	4	4	7707	6234	9.6	11.5	2.30	2.65	.72	.25	.40	.46	.63
Ambition	3	4	7751	6239	12.4	16.4	1.63	2.18	-.06	.35	.47	.49	.67

^aEffect Size = $\frac{\text{Mean}_{LV} - ([\theta \text{ Items}_{LV} / \theta \text{ Items}_{CV}] \times \text{Mean}_{CV})}{\text{Standard Deviation}_{LV}}$. Positive effect sizes indicate higher mean scores in the LV sample.

^bEach item was correlated with the sum of the remaining items in the scale.

Table 3.88**Longitudinal Validation: JOB Score Means and Effect Sizes by Gender**

JOB Scale	Male		Female		Effect Size ^a (d)
	Mean	SD	Mean	SD	
Job Pride	44.0	4.03	44.9	3.83	-.22
Job Security/Comfort	27.0	2.42	27.9	2.17	-.39
Serving Others	11.9	2.01	13.0	1.84	-.54
Job Autonomy	14.5	2.39	14.3	2.55	.11
Routine	11.5	2.66	11.7	2.59	-.07
Ambition	16.4	2.17	16.2	2.24	.13

Note. N for males = 5401-5485; N for females = 827-837.

^ad is the standardized mean difference between male and female scores. A positive value indicates superior performance by males; a negative value indicates superior performance by females.

The means and standard deviations for these scales by race are shown in Table 3.89. Race differences tend to be quite small, with blacks scoring slightly higher than whites on five of the six scales. Hispanics score higher than whites on four of the scales; other minorities score higher than whites on five of the scales.

Uniqueness Analyses

Table 3.90 presents squared multiple regression coefficients, reliability coefficients, and ASVAB uniqueness estimates for JOB scale scores. Only one of the JOB scales--Job Routine--overlaps substantially with the ASVAB, this in spite of the relatively low internal consistency reliability for this scale. In general, the JOB has good potential for building upon the predictive validity of the ASVAB.

Formation of JOB Composites

In an effort to identify clusters of JOB scales that cohere empirically and rationally, we conducted a series of principal components and principal factor analyses on the JOB. The correlations among scale scores for the CV and LV samples are shown in Table 3.91.

Table 3.89

Longitudinal Validation: JOB Score Means and Effect Sizes by Race

JOB Scale	White (N = 4423-1480)		Black (N = 1480-1514)		Effect Size ^a (d)	Hispanic (N = 210-214)		Effect Size ^a (d)	Other (N = 216-217)		Effect Size ^a (d)
	Mean	SD	Mean	SD		Mean	SD		Mean	SD	
Job Pride	44.0	4.05	44.6	3.85	-.15	43.3	4.09	.17	43.6	3.94	.10
Job Security/Comfort	26.9	2.43	27.7	2.24	-.34	27.0	2.49	-.04	27.2	2.37	-.12
Serving Others	12.0	2.05	12.4	1.93	-.20	12.2	1.90	-.10	12.2	2.04	-.10
Job Autonomy	14.5	2.40	14.5	2.47	.00	14.5	2.37	.00	14.6	2.45	-.04
Job Routine	11.2	2.56	12.3	2.72	-.42	11.6	2.74	-.16	11.5	2.65	-.12
Ambition	16.3	2.19	16.8	2.06	-.23	16.7	2.24	-.18	16.6	2.21	-.14

^ad is the standardized mean difference between two subgroups' scores. All effect sizes in this table are relative to the white subgroup. A positive effect size indicates that whites score higher than the minority, and a negative value indicates that whites score lower.

A comparison of the pattern of intercorrelations of JOB scales within the CV sample and the pattern of intercorrelations within the LV sample reveals a different pattern of correlations for the Job Routine scale. This different pattern is likely due to the changes in the Job Routine scale that was administered to the LV sample as part of the Experimental Battery. We described earlier in this chapter the extensive revisions to the JOB Trial Battery inventory. As shown in Table 3.90, the Job Routine scale was more internally consistent in the Experimental Battery version than in prior versions. The changes probably resulted in the different pattern of correlations for the Job Routine scale shown in Table 3.91.

The results of a three-component factor analysis of the JOB scales, with the CV and LV results shown in parallel, are presented in Table 3.92. The CV and LV factor structures are extremely similar. There appears to be one main factor composed of Job Pride, Job Security/Comfort, Serving Others, and Ambition. Two individual scales appear to measure their own unique constructs: Job Routine and Job Autonomy. Consistent with this finding, neither Job Routine nor Job Autonomy correlate highly with any of the other JOB scales. Given the similarity of the CV and LV factor structure, the same composite formation strategy is recommended for the LV data as was used for the CV data. These composites are shown in Figure 3.8.

Table 3.90

Longitudinal Validation: Reliability Coefficients, Multiple Regression Coefficients, and Uniqueness Estimates for JOB Scale Scores: CV/LV

JOB Scale	No. Items	Alpha Coefficient ^a	ASVAB R-Squared ^b	Uniqueness ^c
Job Pride	10/10	.84/.79	.01/.01	.83/.78
Job Security/Comfort	5/6	.67/.76	.01/.02	.66/.74
Serving Others	3/3	.66/.80	.01/.02	.65/.78
Job Autonomy	4/4	.50/.59	.03/.01	.47/.58
Job Routine	4/4	.46/.63	.06/.12	.40/.51
Ambition	3/4	.49/.67	.01/.00	.48/.67

^aN = 7724/6228.

^bR² is adjusted for shrinkage (i.e., cross-validity estimated); N = 6434/4930.

^cAlpha reliability minus ASVAB adjusted R².

Table 3.91

Comparison of JOB Scale Intercorrelations for Revised Trial and Experimental Batteries: CV/LV

JOB Scale	Job Pride	Job Security/Comfort	Serving Others	Job Autonomy	Job Routine
Job Security/Comfort	.61/.65				
Serving Others	.45/.39	.45/.41			
Job Autonomy	.23/.27	.20/.22	.20/.19		
Job Routine	-.24/.07	-.14/.08	-.07/.11	-.09/.08	
Ambition	.50/.50	.45/.42	.33/.28	.20/.28	-.17/-.01

Note. N = 7640-7814/6149-6309.

Table 3.92

Comparison of JOB Principal Components Analysis^a Results for Revised Trial (CV) and Experimental^b Batteries: CV/LV

JOB Scale	High Job Expectations	Job Routine ^c	Job Autonomy ^d	h^2 ^e
Job Pride	.82/.83	-.05/-.01	.06/.18	.68/.73
Job Security/Comfort	.81/.84	-.22/.04	.09/.07	.72/.71
Serving Others	.72/.68	.15/.21	.12/-.01	.56/.50
Ambition	.70/.63	-.19/-.21	.09/.37	.53/.58
Job Routine	-.10/.06	.97/.96	-.04/.05	.95/.94
Job Autonomy	.15/.13	-.04/.08	.99/.96	1.00/.94
Eigenvalue	2.34/2.28	1.02/1.08	1.08/1.02	4.44/4.38

Note: N = 7595/6149.

^aInitial Longitudinal sample (screened).

^bVarimax rotation.

^cThird LV factor.

^dSecond LV factor.

^e h^2 = communality (sum of squared factor loadings) for variables.

Scale	Composite
Pride Job Security Serving Others Ambition	High Job Expectations
Routine	Job Routine
Autonomy	Job Autonomy

Figure 3.8. Longitudinal Validation: Model for formation of JOB composites.

Composite scores were computed by standardizing the component scales to have a mean of 50 and standard deviation of 10, then summing scores for each composite. We then generated split-half reliability estimates for each of the JOB composites. For the first composite, which contains four scales, the scales were first split in half and the halves were added to form composite halves. The Spearman-Brown corrected split-half reliabilities for High Expectations, Routine, and Autonomy are .85, .65, and .47, respectively. The reliability of the first two composites is quite acceptable and higher than the reliability of the third composite (Job Autonomy), which contains only one scale composed of four items.

The correlations among the JOB composite scores are shown in Table 3.93. The correlations are quite low, indicating that they measure relatively independent constructs. The reliability and ASVAB Uniqueness estimates for these composite scores are shown in Table 3.94. The first composite score, High Expectations, appears to measure substantially more unique and reliable variance than is measured by the other two composites. The Routine composite score overlaps the most with ASVAB scores and the Autonomy composite is the least reliable of the three composites.

Table 3.93

Longitudinal Validation: Correlations Among JOB Composite Scores

	High Expectations	Routine
Routine	.09	
Autonomy	.31	.08

Note: N = 6116-6234.

Table 3.94

Longitudinal Validation: JOB Composite Score Reliability and Uniqueness

Composite	Split-Half Reliability ^a	ASVAB R-squared ^b	Uniqueness ^c
High Expectations	.84	.02	.82
Routine	.65	.12	.53
Autonomy	.47	.01	.46

^aSpearman-Brown corrected.

^bR² is adjusted for shrinkage; N = 6134-6167.

^cSplit-half reliability minus ASVAB adjusted R².

The means and standard deviations of the components are shown in Table 3.95 by gender and in Table 3.96 by race. As expected, these composite scores will not adversely impact women or minorities. On the average, women scored higher than men on two of the three composites, and minorities scored higher than whites on all of the composites.

Table 3.95

Longitudinal Validation: JOB Composite Score Means and Effect Sizes by Gender

Composite	Male		Female		Effect Size ^a (d)
	Mean	SD	Mean	SD	
High Expectations	198.7	30.46	208.9	28.57	-.33
Routine	49.9	10.03	50.6	9.77	-.07
Autonomy	50.1	9.90	49.0	10.58	.11

Note. N for males = 5373-5401; N for females = 827-833.

^ad is the standardized mean difference between male and female scores.

A positive value indicates higher scores for males; a negative value indicates higher scores for females.

Table 3.96

Longitudinal Validation: JOB Composite Score Means and Effect Sizes by Race

Composite	White (N = 4308-4326)		Black (N = 1469-1484)		Effect Size ^a (d)	Hispanic (N = 210-213)		Effect Size ^a (d)	Other (N = 204-202)		Effect Size ^a (d)
	Mean	SD	Mean	SD		Mean	SD		Mean	SD	
High Expectations	197.7	30.78	207.2	28.08	-.32	200.0	31.85	-.07	200.7	29.36	-.10
Routine	48.9	9.66	53.1	10.29	-.43	50.6	10.35	-.17	50.2	10.01	-.13
Autonomy	49.9	9.93	50.2	10.23	-.03	50.1	9.83	-.03	50.5	10.16	-.06

^ad is the standardized mean difference between two subgroups' scores. All effect sizes in this table are relative to the white subgroup. A positive effect size indicates that whites score higher than the minority, and a negative value indicates that whites score lower.

LONGITUDINAL VALIDATION PREDICTORS: SUMMARY OF DATA ANALYSES AND FORMATION OF COMPOSITES

As noted in the Introduction to this chapter, there were four major phases in the analysis of the Experimental Battery: screening data, forming basic scores, computing descriptive statistics and conducting psychometric analyses, and developing recommendations for composite scores. Each of these is summarized below.

Data Screening and Basic Scoring

Well-formulated and pilot-tested data collection instruments and procedures are important for ensuring data quality. All instruments had been administered to several sequential samples over the course of earlier phases of Project A, allowing successive refinement of instruments and administration procedures. The intention of the LV data screening was to identify the relatively few remaining problems.

In general, inspections of the data were aimed at identifying two types of undesirable data sets: excessive amounts of missing data and "suspect" data, or test responses that may have been made carelessly, inattentively, or in an otherwise uninformative manner. The predictor instruments differ in their susceptibility to these sources. The computer-administered measures have relatively small amounts of missing data since the presentation of items and recording of responses is under automated control, whereas paper-and-pencil instruments may have relatively more missing data since the examinee is free to skip items. The spatial, paper-and-pencil tests all have time limits, whereas the temperament/biodata, interest, and job preference inventories do not--thus a missing response means different things with regard to these instruments. Different instruments present different data screening problems and different strategies have been used for the various instruments.

For the six spatial tests, several methods of screening data were investigated because it is difficult to differentiate "might-be-random responders" from low-ability examinees. However, so few examinees would have been screened out for four of the six tests, even if there had been no confusion with low-ability examinees, that no special screening was applied to any of the six tests. If an examinee had at least one response, he or she was included and all items were used in computing of scores. Thus, only examinees who had absolutely no data for a test were screened out. With regard to basic scoring, number correct, number wrong, and an accuracy score were compared. We concluded that the number correct score was the most appropriate score to carry forward into the substantive analyses. It appeared to have the best psychometric properties overall and was the method most consistent with the test administration instructions.

For computer-administered tests, very few examinee/test cases were eliminated with the initial screening criteria--ranging from less than one-half of one percent for some psychomotor tests to about 4 percent for the Target Shoot test. Overall, 94 percent of the samples analyzed had complete data for all computer-administered tests. After these minimum data screens, computing of the basic scores for each test involved further minimum data requirements, such as a minimum number of items within each distinct type of item for a given test (e.g., at least three two-character items on the

Perceptual Speed and Accuracy Test). For some tests, several basic scores were compared, such as medians, clipped means, and means of means. Score distributions and reliabilities for the competing scores were compared and an attempt was made to be consistent across similar types of tests. In the end, 17 basic scores were selected. They included simple means, means of means, and medians, depending on the method that appeared to provide the best psychometric properties and provide consistency across very similar tests.

For the ABLE, AVOICE, and JOB, data screening techniques shared a common minimum data screen. That is, an examinee had to have responded to 90 percent of the items on an instrument or the examinee was not scored for that instrument. For the ABLE, the Non-Random Response Scale was applied as an additional screen. This specially developed scale had worked well in prior phases of research and did so again for the Longitudinal Validation data. About 9 percent of ABLE examinees were screened out using these two screening methods, which was slightly less than the percentage screened out for the Concurrent Validation sample.

Four new data-screening techniques were developed for the AVOICE and JOB. In prior samples, we had relied on the ABLE Non-Random Response screen for these two instruments. However, that method assumed that persons not responding attentively to the ABLE were responding in a similar way to the AVOICE and JOB. It seemed preferable to make such screens directly on each instrument, if possible. The four new techniques were different methods of detecting careless or low-literacy examinees. Examinees were not scored for an instrument if they fell above cut scores that were set at the extremes of the distributions of scores on the techniques. The cut scores were set extremely conservatively so that there could be little doubt that an examinee's pattern of responses was not the pattern expected of an attentive, minimally literate examinee. Approximately 10 percent of the examinees were eliminated using these techniques, about the same as the percentage screened by the ABLE Non-Random Response scale in the Concurrent Validation sample.

Keys for summing items to form scale scores were already in existence for the ABLE, AVOICE, and JOB, based on the theoretical and empirical work completed in earlier stages of Project A and Career Force research. These keys were used for these data as well. For the ABLE, we checked the key by correlating each item with its keyed scale score and all the other scale scores. A few items (13 of 199) correlated as much as .05 higher with a non-keyed scale, but not enough to warrant revising the present key.

In general, the data screening techniques and methods of forming basic scores mirrored the procedures used in prior research phases as closely as possible. However, changes were made where it seemed fairly clear that some improvement in psychometric properties could be made or some increase in external validity might be expected.

Descriptive Statistics and Psychometric Properties

With regard to distributional and psychometric properties, a primary concern was to compare the Longitudinal Validation sample results to the Concurrent Validation sample results. While some changes were made to some of the instruments between CV and LV, we expected the psychometric properties to be similar.

There were some score distribution differences (especially for ABLE scales, where most scale scores were elevated in the Longitudinal Validation sample, and a few AVOICE scales that showed much higher mean scores in the Longitudinal Validation sample), but generally they were not large. For most test/scale scores, the variances were very similar in value--indeed the general trend seemed to be slightly greater variance in the Concurrent Validation sample than in the Longitudinal Validation sample. Apparently, the effects of attrition over the course of the first term in the Army did not result in reduced variance of the Concurrent Validation scores as compared to Longitudinal Validation sample scores. However, it must be kept in mind that the Concurrent Validation and Longitudinal Validation samples are different cohorts, and other factors could have operated to keep the variances so similar.

The reliability coefficients and score intercorrelations were remarkably similar across the two samples. Some test/scale scores increased in reliability because of instrument revisions and modifications in scoring methods. Bivariate correlations produced some differences across the samples, but factor analyses showed highly similar solutions. Consequently, the uniqueness (from ASVAB) coefficients were also similar across the two samples. As before, all the Experimental Battery measures showed substantial uniqueness, with the ABLE and JOB scales showing the most, followed fairly closely by the AVOICE; the computer-administered and spatial tests showed relatively more overlap with the ASVAB.

Subgroup differences followed a similar pattern that generally replicated the earlier Concurrent Validation results. The computer-administered and spatial tests recorded the greatest subgroup differences, followed by the AVOICE scales, with the ABLE and the JOB showing the smallest subgroup differences. In the temperament/interest domain, differences often favored minority subgroups. Subgroup differences for Hispanic and Other (minorities) compared to white did show some fluctuations across samples, due to their smaller sample sizes.

Formation of Composite Scores

The basic score analyses produced a set of 72 scores. This number is too large for general validation analyses involving techniques that take advantage of idiosyncratic sample characteristics, such as ordinary least squares multiple regression. Therefore, a series of analyses was conducted to determine an appropriate set of composite scores that would preserve the heterogeneity of the full set of basic scores to the greatest extent possible. These analyses included exploratory factor analyses and confirmatory factor analyses guided by considerable prior theory and empirical evidence (Peterson, et al., 1990, McHenry, et al., 1990). A final set of 31 composites was identified and is shown in Figure 3.9.

Internal consistency and uniqueness estimates and subgroup differences for the composite scores are summarized in Table 3.97, and the intercorrelation matrix of the 31 scores is shown in Table 3.98. Uniqueness, in Table 3.97, provides an estimate of the amount of reliable, unique variance that each composite may contribute to future prediction algorithms. The uniqueness estimates for the temperament/biodata composites from the ABLE and the vocational interest composites from the AVOICE are fairly uniform and

ASVAB Composites	Computer-Administered Test Composites*	ABLE Composites	AVOICE Composites
Quantitative Math Knowledge Arithmetic Reasoning	Psychomotor Target Tracking 1 Distance Target Tracking 2 Distance Cannon Shoot Time Score Target Shoot Distance	Achievement Orientation Self-Esteem Work Orientation Energy Level	Rugged/Outdoors Combat Rugged Individualism Firearms Enthusiast
Technical Auto/Shop Mechanical Comprehension Electronics Information	Movement Time Pooled Movement Time	Leadership Potential Dominance	Audiovisual Arts Drafting Audiographics Aesthetics
Speed Coding Speed Number Operations	Perceptual Speed Perceptual Speed & Accuracy (DT) Target Identification (DT)	Dependability Traditional Values Conscientiousness Modeling	Interpersonal Medical Services Leadership/Guidance
Verbal Word Knowledge Paragraph Comprehension General Science	Basic Speed Simple Reaction Time (DT) Choice Reaction Time (DT)	Adjustment Emotional Stability	Skilled/Technical Science/Chemical Computers Mathematics Electronic Communication
Paper-and-Pencil Test Composite	Perceptual Accuracy Perceptual Speed & Accuracy (PC) Target Identification (PC)	Cooperativeness Cooperativeness	Administrative Clerical/Administrative Warehousing/Shipping
Spatial Assembling Objects Test Object Rotation Test Maze Test Orientation Test Map Test Reasoning Test	Basic Accuracy Simple Reaction Time (PC) Choice Reaction Time (PC)	Internal Control Internal Control	Food Service Food Service - Professional Food Service - Employee
	Number Speed and Accuracy Number Memory (Operation DT) Number Memory (PC)	Physical Condition Physical Condition	Protective Services Fire Protection Law Enforcement
	Short-Term Memory Short-Term Memory (PC) Short-Term Memory (DT)	JOB Composites High Job Expectations Pride Job Security Serving Others Ambition	Structural/Machines Mechanics Heavy Construction Electronics Vehicle Operator

*DT = Decision Time and PC = Proportion Correct

Figure 3.9. Longitudinal Validation Experimental Battery: Composite scores and constituent basic scores.

Table 3.97

**Experimental Battery Scores for Longitudinal Validation Initial Sample:
Reliabilities, Uniqueness Estimates, and Effect Size**

Composite	Internal Consistency ^a	ASVAB Uniqueness ^b	Gender Difference Effect Size ^c	White/ Black Effect Size ^d
<i>ASVAB</i>				
Quantitative	.93	.46	.15	.19
Technical	.93	.43	1.06	1.45
Speed	.88	.74	-.63	-.03
Verbal	.96	.47	.15	1.13
<i>Paper-and-Pencil</i>				
Spatial	.96	.47	.27	1.22
<i>Computer-Administered</i>				
Psychomotor	.94	.75	1.42	.71
Movement Time	.97	.93	.51	.31
Perceptual Speed	.98	.90	.37	.40
Basic Speed	.92	.88	-.01	-.02
Perceptual Accuracy	.75	.72	-.21	.22
Basic Accuracy	.62	.61	-.13	.19
Number Speed & Accuracy	.83	.44	.18	.50
Short-Term Memory	.80	.78	-.05	.19
<i>ABLE</i>				
Achievement Orientation	.91	.89	-.03	-.22
Leadership Potential	.83	.81	.16	-.26
Dependability	.83	.82	-.24	-.23
Adjustment	.81	.79	.18	-.21
Cooperativeness	.68	.67	-.12	-.27
Internal Control	.70	.68	-.24	.08
Physical Condition	.80	.79	.54	-.23
<i>AVOICE</i>				
Rugged/Outdoors	.91	.75	1.13	.67
Audiovisual Arts	.85	.83	-.26	-.35
Interpersonal	.93	.89	-.39	-.45
Skilled/Technical	.92	.87	.00	-.55
Administrative	.93	.84	-.40	-.82
Food Service	.91	.87	-.16	-.52
Protective Services	.85	.83	.36	.23
Structural/Machines	.93	.79	.85	-.11
<i>JOB</i>				
High Job Expectations	.84	.82	-.33	-.32
Job Routine	.65	.53	-.07	-.43
Job Autonomy	.47	.46	.11	-.03

^aInternal consistency estimates for ABLE, AVOICE, JOB, and Computer-Administered Test composites are Spearman-Brown corrected, split-half estimates. The internal consistency for the Spatial and ASVAB composite was estimated using Nunnally's (1978) formula for the reliability of a composite. For the two most speeded spatial tests (i.e., Object Rotation, Maze) separately-timed-halves, corrected correlations were used as reliability estimates in the Nunnally formula; split-half, corrected correlations were used for the remaining four Spatial tests. ASVAB reliabilities were taken from Kass, et al. (1982).

^bInternal consistency estimate minus ASVAB adjusted R^2 . For ASVAB composites, the constituent subtests were excluded from the predictor set.

^cStandardized mean difference between male and female scores. A positive value indicates higher scores for males, and a negative value indicates higher scores for females.

^dStandardized mean difference between scores for blacks and whites. A positive effect size indicates that whites score higher than blacks, and a negative value indicates that whites score lower than blacks.

Table 3.98

Correlations Between Experimental Battery Composite Scores for Longitudinal Validation Initial Sample

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Spatial	.44														
2. ASWAB Verbal	.58	.55													
3. ASWAB Quantitative	.17	-.01	.22												
4. ASWAB Speed	.55	.63	.50	.10											
5. ASWAB Technical	.43	.22	.25	.05	.38										
6. Psychomotor	.34	.17	.19	.17	.15	.31									
7. Perceptual Speed	.17	.12	.12	.06	.10	.07	.43								
8. Perceptual Accuracy	.37	.30	.57	.32	.24	.22	.21	.10							
9. Number Speed/Accuracy	.10	-.01	.07	.18	.06	.17	.31	.12	.15						
10. Basic Speed	.09	.06	.08	.04	.05	.03	.02	.09	.08	.01					
11. Basic Accuracy	.27	.08	.17	.22	.05	.21	.34	.03	.25	.27	.09				
12. Short-Term Memory	.20	.06	.08	.10	.14	.31	.24	.09	.09	.10	.09	.18			
13. Movement Time	-.09	-.02	-.05	.08	.08	.07	.03	.00	.04	.02	-.02	.00	.03		
14. High Job Expectations	-.20	-.33	-.24	-.01	-.25	.13	.08	.03	.13	.02	-.02	-.05	-.06	.08	
15. Job Routine	.02	.08	.05	-.02	.11	.01	.00	.02	.01	.02	-.01	.04	.10	.30	.07
16. Job Autonomy	.07	.06	.10	.10	.08	.07	.03	.00	.05	.04	-.02	.05	.10	.35	.10
17. Achievement Potential	.05	.10	.09	.08	.08	.08	.05	.01	.04	.05	-.01	.04	.08	.35	.16
18. Leadership Potential	.01	.01	.03	.08	.01	.01	.01	.07	.00	.01	.02	.06	.04	.29	.08
19. Dependability	.09	.09	.10	.05	.11	.10	.04	.00	.05	.04	.02	.05	.08	.20	.12
20. Adjustment	.03	.04	.04	.06	.03	.05	.01	.05	.02	.02	.01	.05	.04	.28	.06
21. Cooperativeness	.11	.12	.10	.07	.07	.05	.02	.05	.07	.01	.03	.07	.12	.14	.03
22. Internal Control	-.01	-.06	.01	.04	-.02	.14	.09	.09	.04	.10	.07	.03	.12	.14	.03
23. Physical Conditions	.20	.18	.08	.11	.36	.32	.14	.02	.02	.01	.00	.01	.17	.10	.03
24. Rugged/Outdoors	.03	.07	.04	.03	.00	.00	.00	.04	.05	.01	.01	.02	.02	.28	.01
25. Audiovisual Arts	-.06	.03	.03	.10	.10	.07	.02	.01	.00	.03	.00	.04	.03	.48	.01
26. Interpersonal	.02	-.07	.12	.11	.03	.00	.03	.04	.06	.02	.00	.06	.02	.26	.08
27. Skilled/Technical	-.16	-.22	-.10	.10	.24	.19	.14	.03	.06	.02	.00	.01	.08	.23	.29
28. Administrative	-.11	-.13	.06	.03	.15	.13	.10	.01	.06	.04	.01	-.04	-.06	.07	.23
29. Food Service	.04	.02	-.03	-.01	.08	.11	.08	.04	.01	.03	.06	.04	.08	.20	.05
30. Protective Services	.02	-.12	-.06	-.10	.18	.11	.04	.01	.04	.04	.01	-.03	.06	.06	.17
31. Structural/Machines															

(cont inued)

Table 3.98 (Continued)

Correlations Between Experimental Battery Composite Scores for Longitudinal Validation Initial Sample

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
16. Job Autonomy	--															
17. Achievement Orientation	.22	--														
18. Leadership Potential	.24	.66	--													
19. Dependability	.07	.61	.33	--												
20. Adjustment	.14	.70	.53	.40	--											
21. Cooperativeness	.07	.59	.39	.60	.53	--										
22. Internal Control	.07	.54	.33	.51	.44	.45	--									
23. Physical Conditions	.09	.49	.35	.22	.36	.22	.18	--								
24. Rugged/Outdoors	.18	.22	.18	.07	.20	.09	.13	.19	--							
25. Audiotvisual Arts	.20	.23	.21	.19	.15	.21	.13	.07	.17	--						
26. Interpersonal	.18	.41	.47	.30	.29	.31	.25	.19	.15	.56	--					
27. Skilled/Technical	.19	.30	.25	.25	.21	.21	.18	.11	.21	.58	.55	--				
28. Administrative	.07	.14	.12	.21	.06	.13	.07	.01	.01	.45	.47	.58	--			
29. Food Service	.03	.03	.04	.06	.00	.04	.01	-.02	.01	.30	.29	.26	.21	--		
30. Protective Services	.15	.22	.23	.15	.15	.16	.13	.14	.52	.19	.38	.24	.21	.14	--	
31. Structural/Machines	.15	.09	.04	.01	.06	.05	.01	.11	.52	.28	.14	.42	.37	.29	.37	--

Note. N = 4623.

moderately high, ranging from .67 to .89 for the ABLE composites and from .75 to .89 for the AVOICE composites. Uniqueness estimates for the JOB composites are somewhat lower, partly because the composites are less homogeneous. There is much greater variability in the uniqueness estimates for the cognitive, perceptual, and psychomotor composites. Movement Time and Perceptual Speed composites have large amounts of new information to add to such prediction problems. The Spatial and the Number Speed and Accuracy composites contribute less unique information, but they still have respectable levels of additional, reliable variance to contribute.

Past validation analyses (McHenry, et al., 1990) have shown that all of these measures have reasonably high concurrent relationships with the job performance of first-term Army enlisted personnel, and that these relationships vary across the various job performance criterion measures. Also, the ABLE, and to a somewhat lesser extent the AVOICE, show considerably smaller subgroup differences than do the ASVAB, Spatial, and some of the computer-administered composites. This would seem to be of some importance in forming equations that would show effective prediction with the least possible differential impact across subgroups.

A FINAL WORD

This chapter has focused on analyses of the Experimental Battery measures that were administered to the Longitudinal Validation sample. Comparisons with earlier versions of the battery (the Trial Battery and Revised Trial Battery) revealed some differences, but no major dissimilarities or inconsistencies. Procedures were developed for screening data and producing psychometrically sound scale/test scores. Further analyses identified a set of composite scores that will be used in the validation analyses to be conducted in the second and third years of the project.

Chapter 4

End-of-Training Measures: LV Sample

This chapter describes preliminary analyses of the results of the Longitudinal Validation End-of-Training (EOT) measures: the EOT written School Knowledge (SK) tests and the EOT rating scales. The primary purpose of these analyses was to aggregate the individual test items and rating scales into a set of EOT composite scores that would be consistent with the model of first-tour performance.

The EOT SK tests are paper-and-pencil achievement tests that were designed to assess the knowledge possessed by soldiers after they have completed MOS-specific Advanced Individual Training. These tests contain between 97 and 180 multiple-choice items measuring both technical knowledge specific to a MOS and more general, Army-wide knowledge relevant to all MOS.

The seven EOT rating scales were modified versions of a subset of the Army-wide Behaviorally Anchored Rating Scales (BARS) used as job performance measures in the Concurrent Validation phase of Project A. Ratings were obtained from supervisors (drill instructors) as well as peers (classmates).

THE END-OF-TRAINING (EOT) SCHOOL KNOWLEDGE (SK) TESTS

A detailed description of the development and field testing of the School Knowledge tests is provided in the FY 85 Project A Annual Report (Campbell, 1987). Briefly, an initial item pool was developed and subsequently reviewed both by job incumbents and by school trainers. Items were then administered to trainees to test for clarity. Based upon the comments obtained from the reviews and administration, items were revised, and then field tested by administering them to job incumbents. After further reviews by Army Training and Doctrine Command (TRADOC) proponent agencies, the items were revised for administration during the Concurrent Validation phase of Project A. The tests were again sent to TRADOC for proponent review and updating before their administration as End-of-Training tests during the Longitudinal Validation. The tests were administered to both the nine Batch A MOS (the MOS for which job knowledge and hands-on tests were developed) and the 10 Batch Z MOS (the MOS having no job knowledge or hands-on tests).

Basic composite scores for the EOT SK tests were developed in two phases. In the first phase, the test scoring keys were thoroughly reviewed and revised based upon suggestions from proponents, subject matter experts, and results from item analyses. In the second phase, the revised tests were analyzed, using principal components and confirmatory factor analyses. The results of these analyses guided the creation of the basic scores.

Revision of the EOT SK Test Items and Scoring Keys

The first requirement for creating EOT SK test scores was to develop a standard scoring procedure for the tests themselves. The procedure adopted had three steps: (a) compare various scoring keys with the tests to ensure that all keys were identical with the items, (b) examine item parameters generated by an item analysis program to help identify problematic test items,

and (c) review the problematic items targeted by the item analyses to check for keying errors or content/policy changes that would suggest rekeying the item.

Comparison of Scoring Keys

Several sets of SK test scoring keys were available at various stages of test development. These keys needed to be organized and tracked so that only the most current keys would be used to score the test items for the item analyses. The original SK test scoring keys were altered after each test received proponent review in which subject matter experts examined the SK test for their MOS to ensure that item content reflected current Army policy. This review resulted in items from some tests being rescored or dropped.

The final stage involved three steps: (a) The test booklets containing the keyed answers to the items were compared with the original keys. Any discrepancies were noted and rectified. (b) The resulting altered SK test keys were compared with the information obtained from the proponent reviews, to ensure that all requested, authorized changes in the scoring of the items due to content changes had been made. These corrected keys were compared to the keys used to score the items in the item analysis program. A few discrepancies were identified; past records and proponent review sheets were used to select the best keying for all variously keyed items. (c) Finally, the amended SK keys used to score items in the revised item analysis program were compared to the printouts generated by the program to ensure that all desired changes in item keyings had been incorporated.

This extensive review process resulted in a standard set of SK test scoring keys that was believed to reflect the most current Army policy for each MOS.

Item Analysis

After the scoring keys had been examined, the next step was to examine each item from each SK test. An item analysis program provided such item parameters as the keyed response; the proportion of individuals who endorsed each item response option; the mean total score of those individuals who endorsed a particular item response option, excluding the item in question; and the point-biserial correlation between an item and the total score on the test, excluding the item in question.

This program also provided various item flags that earmarked potentially undesirable item parameters, such as an infrequently endorsed distractor, a poorly discriminating item (signified by a point-biserial correlation below .20), and a potentially miskeyed item (signified by a positive point-biserial correlation for one of the distractors). Purposely liberal parameter values were established as cutoffs for the flags so that all items with somewhat suspicious characteristics would be carefully examined. For all flagged items, representing about 40 percent of the total item pool, the item parameters were examined.

Although we wished to identify all items with potentially unusual parameters, we also wished to retain as many items as possible. Thus, based

upon this preliminary examination, the following restrictive item retention/deletion decision rules were established:

- Items flagged only as potentially mis-keyed were retained after the correct answers to the items had been identified. Virtually every item of this type had a distractor characterized by a point-biserial correlation that, although positive, was essentially zero.
- Items flagged only as poorly discriminating were retained. Virtually every item of this type barely made the .20 cutoff, and their retention preserved content coverage.
- Items receiving both types of flags were candidates for deletion from scoring.

Review of Problematic Items

For the items receiving both mis-keying and low discrimination flags, additional indicators were reviewed. Items characterized by all of the following item parameters were labeled "problematic":

- Higher endorsement rates for a distractor than for the keyed response.
- Higher discrimination indexes for a distractor than for the keyed response.
- Higher mean scores on all items, excluding the item in question, for those who endorsed the distractor than for those who responded correctly to the item.

A total of 172 items were identified by all three screens. Ten other items that were flagged by two of the three screens were also included as borderline cases because the parameters for which they were flagged exceeded the cutoff values by a wide margin. Thus, 182 items were labeled problematic, a figure that represents 5.6 percent of the total pool of 3,312 SK test items.

Although the review of the scoring keys resulted in perfect correspondence between all the keys and the scoring of the item analysis program, the parameters for several of the items continued to suggest keying errors, implying that the latest versions of the SK test keys received from the proponents were in error. To identify and correct these discrepancies, a booklet containing the 182 problematic items and their respective item statistics was assembled and disseminated to various project staff. This exercise permitted a final perusal of the items by subject matter experts and project staff in an effort to identify any keying errors, or changes in training content or Army policy. This review of the item parameters also allowed modifications to be made to the strategy used to identify problematic test items.

In addition to the item booklet review, further evidence on deleting problematic items was obtained by examining the parameters of the SK test items for the LVI sample. This comparison was made for the Batch Z MOS only,

because the Batch A MOS were administered job knowledge tests during the LVI data collection. Items having undesirable item parameters in only one of the samples were retained. Those items having undesirable item parameters in both samples were deleted from the tests.

The LVI revisions and deletions of the EOT SK test items are summarized in Table 4.1. Information is provided on the total number of items on the original SK tests, the number of items rescored, the number of items included in the test but not scored (mostly because the tests were administered before the proponent reviews that suggested these items be dropped had been returned, and occasionally because of a reliability problem, so that these items had not been screened as described above), the number of items dropped, and the total number of items used to calculate the revised total SK test score. A total of 37 items were rescored and 17 items were dropped as a result of the item review procedures. Thus, 165 of the 182 problematic items were retained in some fashion.

Analyses of the Revised EOT SK Tests

The next steps in the analysis of the revised tests were (a) screening of the EOT SK tests for missing and random data, (b) calculation of functional category scores, (c) principal components analyses of the intercorrelations among the functional categories, (d) confirmatory factor analyses of the functional categories, and (5) calculation of the EOT SK test factor scores.

Data Screens

Two screens were employed to address the problems of missing and random data. First, individuals who responded to less than 90 percent of the SK test items were treated as missing. Second, point-biserial correlations were calculated between the score on a particular item (1 if correct and 0 if incorrect), and the difficulty of that item (the proportion of correct responses to an item). These values should be positive, because easy items are correctly answered by most examinees whereas difficult items are correctly answered by fewer examinees. Examinees with point-biserial correlations that were essentially zero were tagged as random responders and were removed from consideration.

In addition to these two data screens, item scores were imputed for missing data internal to the test (i.e., items that the examinees reached but did not answer, although they answered subsequent items). The score given for these items was the lesser of the chance score and the item difficulty. No imputations were calculated for external missing data (i.e., those items not reached at the end of the test, evidenced by failure to answer subsequent items). Imputation of these items will be conducted when the factor scores are computed for future analyses; in the present analyses, these items were treated as missing.

Table 4.1**Summary of Item Alterations for the EOT School Knowledge Tests**

MOS		Original Number of Items	Items Rescored	Items Not Scored	Items Dropped	Remaining Items
11B	Infantryman	143	3	0	1	142
12B	Combat Engineer	155	3	5	0	150
13B/S	Cannon Crewman	160	2	0	2	158
13B/T	Cannon Crewman	161	1	0	2	159
16S	MANPADS Crewman	148	0	0	2	146
19E	M60 Armor Crewman	162	1	0	0	162
19K	M1 Armor Crewman	160	2	0	0	160
27E	TOW/Dragon Repairer	168	1	0	3	165
29E	Electronics Repairer	150	1	2	1	147
31C	Single Channel Radio Operator	162	0	0	1	161
51B	Carpentry/Masonry Specialist	166	7	12	0	154
54B	NBC Specialist	153	1	0	1	152
55B	Ammunition Specialist	180	0	0	0	180
63B	Light-Wheel Vehicle Mechanic	151	1	0	0	151
67N	Utility Helicopter Repairer	160	2	1	3	156
71L	Administrative Specialist	97	0	0	0	97
76Y	Unit Supply Specialist	165	2	0	0	165
88M	Motor Transport Operator	130	2	0	1	129
91A	Medical Specialist	167	1	0	0	167
94B	Food Service Specialist	127	4	12	0	115
95B	Military Police	120	3	0	0	120
96B	Intelligence Analyst	161	0	0	0	161

Functional Categories

Scale scores for the SK tests were obtained by calculating the average score on the items constituting each functional category on the test. The functional categories are based on the content of the items on the test. They are item content categories established as part of the Concurrent Validation sample (CVI) analysis. Each functional category is characterized by some portion of the knowledges, skills, and equipment that are required to demonstrate adequate performance for a given MOS. Kuhn (1987) described the derivation of the functional categories for all of the MOS from the present analysis except for 29E and 96B, which were added to the sample as part of the Longitudinal Validation. Functional categories for these MOS were derived through a rational classification process by the project staff.

Some functional categories are relevant to all MOS, whereas others comprise items that tap content specific to a particular MOS. The functional categories had been condensed into six factors in earlier analyses of the SK tests conducted as part of Project A (e.g., Kuhn, 1987). The six factors were Communication, Vehicles, Basic Skills, Identify Targets, Technical, and Safety/Survival (CVBITS). This six-factor solution served as a focal point for the present analyses. Table 4.2 lists the 13 common categories, three examples of MOS-specific categories, and the relationship of each category to the CVBITS categorization.

Table 4.2

Functional Categories for the EOT School Knowledge Tests Along With Their CVBITS Classification

Common Categories

First Aid (S)
Navigation (B)
NBC (S)
Weapons (B)
Field Techniques (B)
Communications (C)
Antiaircraft/Antitank (Identify Target)(I)
Antiaircraft/Antitank (Engage Target) (B)
Customs and Laws (B)
Drive (V)
Preventive Maintenance (V)
Vehicle Operation/Recovery (V)
Generators (B)

CVBITS

Safety/Survival
Basic Skills
Safety/Survival
Basic Skills
Basic Skills
Communication
Identify Targets
Basic Skills
Basic Skills
Vehicles
Vehicles
Vehicles
Basic Skills

Examples of MOS-Specific Categories

Install Electronic Equipment (T - 31C)	Technical
Forms/Files Management (T - 71L)	Technical
General Medical Knowledge (T - 91A)	Technical

Principal Components Analyses

Once the functional category scores had been calculated for each MOS, they were analyzed via principal components. Solutions containing between two and seven components were retained and subjected to a varimax rotation for each MOS. The results of the components analyses strongly suggested the presence of one or two components--a Basic Skills component and an MOS-specific, or Technical, component. (The one-component solution appeared for such MOS as 11B and 88M, suggesting a Technical component only.) There was little correspondence between the six-component solution and the CVBITS solution.

Confirmatory Factor Analyses (CFA)

The results from the present set of components analyses and the previous analyses from Project A provided three potential models of EOT SK test performance:

- CFA1: A one-factor solution
- CFA2: A two-factor solution (Basic and Technical)
- CFA3: A variable-factor solution (CVBITS)

The CVBITS solution is variable across MOS because the SK tests for some MOS do not tap all of the 13 functional categories. If all of the functional categories within one of the CVBITS factors are excluded from the content of a test, there will be a corresponding reduction in the factor structure defining that test. Thus, for MOS 95B, the third model tested was CVBTS, because this test has no items relating to identification of targets. The CFA3 models tested for each MOS are listed in Table 4.3.

To test empirically which of the three models appeared most plausible for the EOT sample, their relative fit was assessed with confirmatory factor analyses using LISREL (Joreskog & Sarbom, 1986). For models two and three, there are occasions when a functional category classified as Basic for most MOS is considered Technical for a particular MOS due to the importance of that category's content to the MOS in question. For example, the Basic functional category First Aid is considered a Technical category for the Medical Specialists (MOS 91A). Thus, the three models that were tested for each MOS differ in terms of the functional categories that were specified for the latent variables. The sample sizes for each MOS used in the confirmatory analyses are presented in Table 4.4.

In addition to the usual statistical measures that allow assessment of fit in LISREL (e.g., the chi-square value for the model, root mean square residual, goodness-of-fit index, normalized residuals), values for the modification index (MI) were obtained for each model. The MI in LISREL provides an indication of the minimum reduction in the chi-square statistic that could be expected if a targeted parameter were freed. Thus, the reduction in chi-square could be larger than the MI value if the model were relaxed by the freeing of the parameter in question. Additional modified models were tested in the present analyses for some MOS if the MI from the original model suggested that a functional category should load on a different latent variable.

Table 4.3**CVBITS Factors Included in the Confirmatory Factor Analyses of Model CFA3**

MOS	CFA3 Model	MOS	CFA3 Model
11B	CVITS	55B	CVBITS
12B	CVBITS	63B ^a	CBTS
13B ^a	CBTS	67N	CBITS
16S	CVBITS	71L	CVBTS
19E ^a	CBITS	76Y	CVBITS
19K ^a	CBITS	88M ^a	CBITS
27E	CVBITS	91A	VBTS
29E	CVBITS	94B	CVBTS
31C	VBITS	95B	CVBTS
51B	CBITS	96B	CVBITS
54E	CVBITS		

^a Vehicles is subsumed by the Technical category for this MOS.

Table 4.4**Number of Soldiers Included in the Confirmatory Factor Analyses and the Correlational Analyses**

MOS	N for CFAs	MOS	N for CFAs
11B	10575	54E	808
12B	2001	55B	671
13B/S	4364	63B	1451
13B/T	923	67N	408
16S	693	71L	1843
19E	471	76Y	2289
19K	1658	88M	1913
27E	166	91A	5368
29E	306	94B	2693
31C	1376	95B	3776
51B	377	96B	253

Two points should be made about the use of the MI. First, because this index can capitalize on chance fluctuations in the sample, alterations in the model based upon this index should be chosen only if the suggested alterations make sense substantively. Ideally, these alterations should also be cross-validated to ensure that the increase in fit was not limited to the sample. Second, the value of the MI in the present analyses does not reflect the minimum reduction in chi-square that could be expected from model modification, because the parameter having the largest MI was not freed; rather, the parameter was reallocated. If the indicator were allowed to load upon both the original and suggested latent variables, then the MI would retain its usual interpretation. Because no multiple factor loadings were permitted in these analyses, the observed reductions in chi-square in the revised models are less than the quantities suggested by the MI.

The results of the confirmatory factor analyses for all MOS are presented in Table 4.5. This table lists the MOS, the models that were tested for that MOS, the chi-square value and degrees of freedom for each model, the corresponding values for the root mean square residual and goodness-of-fit index for each model, and the MI results.

The MI results in Table 4.5 provide the value of the maximum MI for a particular model, the functional category for which the MI was obtained, and the latent variable upon which the category should load. The revised models (those that employ the MI alteration) are labeled as RCFA2 or RCFA3. Thus, for MOS 91A, the maximum MI for CFA2 is 24.9 and suggests that the First Aid category should load on the Technical factor rather than the Basic factor. This change was incorporated and tested in model RCFA2. The result was a significantly lower chi-square than for CFA2. The MI now suggests that First Aid should load on Basic, although the MI is not as high as for CFA2. Because first aid is of central importance to the job of Medical Specialist, the alteration was deemed sensible. Thus, the revised model was adopted, allowing First Aid to become part of this job's Technical factor.

The three models tested in the present analyses are hierarchical. That is, they are subsets of one another, with the one-factor model (CFA1) being a more general instance of both the two-factor and CVBITS models (CFA2 and CFA3), and the two-factor model a more general instance of CVBITS. This relationship among models allows their relative fit to be determined by comparing their chi-square statistics (for a particular MOS). Specifically, the difference in the chi-square values from any two hierarchical models is distributed as chi-square with degrees of freedom equal to the difference in the degrees of freedom from the two models. If a more specialized model (i.e., a model with more parameters to be estimated, and therefore a less restrictive model) results in a significant decrease in the chi-square value relative to a second model with fewer parameters, then the more specialized model is judged to be a better model because a significant increase in fit has been obtained (i.e., the increase in fit is worth the decrease in the number of degrees of freedom from a statistical standpoint).

Consider the models for MOS 51B in Table 4.5. Model CFA1 yields a chi-square value of 211.5 with 90 degrees of freedom (df). By comparison, CFA2 gives a chi-square of 186.5 with df = 89. Thus, CFA2 reduces the chi-square value by 25.0 and 1 degree of freedom. Because the critical chi-square value with df = 1 is 3.84 (all comparisons based on $p < .05$), and because 25.0 is greater than 3.84, model CFA2 provides a significant increase in fit relative

to CFA1. Similarly, model CFA3 with chi-square of 159.0 and 84 degrees of freedom decreases the chi-square of CFA2 by 27.5 and 5 degrees of freedom. The critical chi-square value for $df = 5$ is 11.07, indicating a significant increase in fit for the CVBITS model (CFA3) over the two-factor model.

The results in Table 4.5 suggest that the two-factor solution (Basic and Technical) is the most stable. Although the CVBITS models resulted in significant increases in fit for 18 of the 22 MOS (MOS 19E, 27E, 29E, and 88M were the exceptions), the increases in fit for eight of these 18 MOS were obtained at the expense of an estimated factor correlation matrix that was not positive definite (e.g., MOS 13B/S, 31C). That is, no allowable factor structure could give rise to the estimated parameter values for these models. Thus, the increase in fit afforded by the CVBITS models is not universal and is often the result of impossible model parameters. In addition, the CFA3 models are not generalizable (i.e., the CVBITS factor structure varies across MOS). Such being the case, the two-factor model was adopted.

Table 4.5

Indexes of Fit Generated by Various Models for the EOT SK Tests

MOS	Model	$\chi^2_{(df)}$	RMS	GFI	Modification Index ^a
11B	CFA1	398.3 ₍₃₅₎	.02	.99	N/A
	CFA3	375.5 ₍₃₁₎	.02	.99	Navigate-->IDENTIFY (17.1)
12B	CFA1	513.4 ₍₁₁₉₎	.05	.96	N/A
	CFA2	423.4 ₍₁₁₈₎	.04	.97	Navigate-->TECH (65.3)
	CFA3	295.1 ₍₁₀₈₎	.03	.97	Navigate-->TECH (63.3)
	RCFA2	383.1 ₍₁₁₈₎	.03	.98	FieldTech-->TECH (35.5)
	RCFA3	251.0 ₍₁₀₈₎	.03	.99	FieldTech-->TECH (30.5)
13B/S	CFA1	224.2 ₍₄₄₎	.02	.99	N/A
	CFA2	210.2 ₍₄₃₎	.02	.99	Weapons-->TECH (45.3)
	CFA3 ^b	156.6 ₍₄₀₎	.02	.99	Weapons-->TECH (23.3)
	RCFA2	159.9 ₍₄₃₎	.02	.99	FirstAid-->TECH (17.1)
	RCFA3	142.5 ₍₄₀₎	.02	.99	Weapons-->SAFETY (10.7)
13B/T	CFA1	69.8 ₍₄₄₎	.03	.99	N/A
	CFA2	69.7 ₍₄₃₎	.03	.99	Weapons-->TECH (17.8)
	CFA3 ^b	63.5 ₍₄₀₎	.03	.99	Weapons-->TECH (5.6)
	RCFA2	57.0 ₍₄₃₎	.02	.99	Howitzer&Ammo-->BASIC (6.2)
16S	CFA1	293.9 ₍₉₀₎	.06	.94	N/A
	CFA2	257.8 ₍₈₉₎	.06	.95	A/A(ID)-->TECH (22.1)
	CFA3	191.2 ₍₇₉₎	.05	.96	Navigate-->TECH (47.0)
	RCFA2	238.0 ₍₈₉₎	.05	.95	Navigate-->TECH (49.0)
	XCFA2 ^c	207.3 ₍₈₉₎	.05	.96	Navigate-->BASIC (18.3)
	RCFA3	141.0 ₍₇₉₎	.03	.97	Navigate-->SAFETY (10.9)
19E	CFA1	114.8 ₍₄₄₎	.05	.96	N/A
	CFA2	99.7 ₍₄₃₎	.05	.96	CustLaw-->TECH (11.5)
	CFA3	90.1 ₍₃₈₎	.05	.97	CustLaw-->TECH (23.9)

Table 4.5 (Continued)

Indexes of Fit Generated by Various Models for the EOT SK Tests

MOS	Model	$\chi^2_{(df)}$	RMS	GFI	Modification Index ^a
19K	CFA1	397.3 ₍₃₅₎	.06	.95	N/A
	CFA2	252.6 ₍₃₄₎	.04	.97	Commun-->TECH (44.5)
	CFA3 ^b	149.5 ₍₂₉₎	.03	.98	Navigate-->TECH (29.8)
	RCFA2	247.2 ₍₃₄₎	.04	.97	Commun-->BASIC (38.8)
27E	CFA1	154.4 ₍₁₀₄₎	.09	.88	N/A
	CFA2	137.6 ₍₁₀₃₎	.08	.90	CustLaw-->TECH (4.6)
	CFA3	121.4 ₍₉₃₎	.07	.91	FirstAid-->SAFETY (8.3)
29E	CFA1	202.1 ₍₉₀₎	.11	.91	N/A
	CFA2	177.7 ₍₈₉₎	.09	.92	Drive-->TECH (32.2)
	CFA3	160.4 ₍₇₉₎	.08	.93	PrevMain-->TECH (21.4)
31C	CFA1	332.5 ₍₉₀₎	.05	.97	N/A
	CFA2	142.5 ₍₈₉₎	.03	.99	A/A(ID)-->TECH (8.3)
	CFA3 ^b	124.7 ₍₈₂₎	.02	.99	Drive-->IDENTIFY (13.5)
51B	CFA1	211.5 ₍₉₀₎	.08	.93	N/A
	CFA2	186.5 ₍₈₉₎	.07	.93	FieldTech-->TECH (8.6)
	CFA3	159.0 ₍₈₄₎	.06	.94	CustLaw-->TECH (12.9)
54E	CFA1	382.8 ₍₇₇₎	.07	.93	N/A
	CFA2	373.9 ₍₇₆₎	.06	.93	NBC-->TECH (47.4)
	CFA3	245.8 ₍₆₆₎	.05	.96	NBC,FirstAid-->VEHICLES (33.4)
	RCFA2	329.4 ₍₇₆₎	.05	.94	Navigate-->TECH (33.1)
55B	CFA1	481.1 ₍₁₀₄₎	.09	.91	N/A
	CFA2	432.7 ₍₁₀₃₎	.08	.92	Ammo-->BASIC (10.1)
	CFA3	348.0 ₍₉₃₎	.08	.93	Navigate-->VEHICLES (26.0)
	RCFA2	472.7 ₍₁₀₃₎	.08	.91	Ammo-->TECH (49.7)
63B	CFA1	530.8 ₍₆₅₎	.05	.94	N/A
	CFA2	484.9 ₍₆₄₎	.04	.95	Drive-->BASIC (120.9)
	CFA3	483.2 ₍₆₁₎	.05	.95	Drive-->BASIC (112.0)
	RCFA2	414.1 ₍₆₄₎	.04	.96	PrevMain-->BASIC (118.8)
	XCFA2 ^c	298.5 ₍₆₄₎	.03	.97	Drive-->TECH (47.5)
	RCFA3	359.2 ₍₆₁₎	.04	.96	PrevMain-->SAFETY (91.0)
67N	CFA1	120.9 ₍₆₅₎	.06	.95	N/A
	CFA2	89.5 ₍₆₄₎	.05	.97	Commun-->TECH (5.8)
	CFA3	75.9 ₍₅₉₎	.05	.97	FirstAid-->COMMUN (5.5)
71L	CFA1	243.8 ₍₆₅₎	.04	.98	N/A
	CFA2	137.7 ₍₆₄₎	.03	.99	FirstAid-->TECH (6.3)
	CFA3	119.1 ₍₅₉₎	.02	.99	CustLaw-->SAFETY (16.7)

(Continued)

Table 4.5 (Continued)

Indexes of Fit Generated by Various Models for the EOT SK Tests

MOS	Model	$\chi^2_{(df)}$	RMS	GFI	Modification Index ^a
76Y	CFA1	912.7 ₍₁₁₉₎	.07	.95	N/A
	CFA2	529.9 ₍₁₁₈₎	.04	.97	Weapons-->TECH (79.4)
	CFA3 ^b	408.2 ₍₁₀₈₎	.03	.98	Weapons-->TECH (71.2)
	RCFA2	701.9 ₍₁₁₈₎	.05	.96	Weapons-->BASIC (248.0)
88M	CFA1	288.5 ₍₄₄₎	.05	.97	N/A
	CFA2	282.3 ₍₄₃₎	.05	.97	FieldTech-->VEHICLES (39.4)
	CFA3 ^b	275.1 ₍₄₀₎	.04	.97	PrevMain-->BASIC (43.7)
91A	CFA1	163.4 ₍₂₇₎	.03	.99	N/A
	CFA2	71.9 ₍₂₆₎	.01	.99	FirstAid-->TECH (24.9)
	CFA3 ^b	61.1 ₍₂₃₎	.01	.99	FirstAid/NBC-->TECH (25.4)
	RCFA2	67.3 ₍₂₆₎	.01	.99	FirstAid-->BASIC (20.3)
	RCFA3	58.0 ₍₂₃₎	.01	.99	FirstAid-->BASIC (22.1)
94B	CFA1	752.5 ₍₆₅₎	.06	.96	N/A
	CFA2	475.7 ₍₆₄₎	.04	.97	FSAdmin-->BASIC (92.6)
	CFA3	322.1 ₍₅₉₎	.03	.98	FSAdmin-->BASIC (65.4)
95B	CFA1	789.2 ₍₆₅₎	.05	.97	N/A
	CFA2	380.8 ₍₆₄₎	.03	.98	Commun-->TECH (64.3)
	CFA3 ^b	275.3 ₍₅₇₎	.03	.99	Navigate-->VEHICLES (67.2)
96B	CFA1	177.2 ₍₉₀₎	.07	.91	N/A
	CFA2	158.8 ₍₈₉₎	.06	.92	Navigate-->TECH (20.4)
	CFA3	132.6 ₍₈₁₎	.06	.93	Navigate-->TECH (16.0)
	RCFA2	143.1 ₍₈₉₎	.05	.92	Navigate-->BASIC (6.5)
	RCFA3	119.7 ₍₈₁₎	.05	.94	Navigate-->BASIC (6.7)

Note. CFA = confirmatory factor analysis. RCFA = Revised CFA. RMS = root mean square residual. GFI = Goodness-of-fit index. MI = modification index.

^aThe obtained reduction in χ^2 is less than the MI given in parentheses.

^cAn additional run taking the RCFA2 MI into account.

^bPsi matrix (correlation matrix of the factors) is not positive definite.

CREATION OF EOT SCHOOL KNOWLEDGE TEST FACTOR SCORES

Using the results from the confirmatory factor analyses, two EOT SK test factor scores were created: a Basic score and a Technical score. For two MOS (11B and 88M), all functional categories are considered Technical. Thus, these MOS have only one of the two factor scores. Each factor score represents the unit-weighted sum of the functional categories that were indicators for these factors in the confirmatory factor analyses. The functional categories and the factor scores to which they contribute are listed in Table 4.6.

Table 4.6**Functional Categories Comprising the Two Subscores for the EOT School Knowledge Tests**

MOS	Basic	Technical
11B	No Basic Categories	First Aid Navigate NBC Weapons Field Techniques Communications A/A ^a (ID) A/A (Engage) Customs and Laws Drive
12B	First Aid NBC Weapons Field Techniques Communications A/A (ID) A/A (Engage) Customs and Laws Drive Preventive Maintenance	Navigate Rigging Field Fortifications and Obstacles Demolitions Construction Tools and Materials Bridging/River Crossing Generators
13B	First Aid Navigate NBC Field Techniques Communications A/A (Engage) Customs and Laws	Weapons Drive Operate Howitzer Sights Prepare, Operate, Maintain Howitzer and Ammunition
16S	First Aid NBC Weapons Field Techniques Communications A/A (Engage) Customs and Laws Drive Preventive Maintenance	Navigate A/A (ID) Redeye Stinger SHORAD IFF

(Continued)

Table 4.6 (Continued)

Functional Categories Comprising the Two Subscores for the EOT School Knowledge Tests

MOS	Basic	Technical
19E	First Aid Navigate NBC Weapons Field Techniques Communications A/A (ID) Customs and Laws	Preventive Maintenance Operate Tanks Tank Gunnery
19K	First Aid Navigate NBC Weapons Field Techniques Communications A/A (ID)	Preventive Maintenance Operate Tanks Tank Gunnery
27E	First Aid Navigate NBC Weapons Field Techniques Communications A/A (ID) A/A (Engage) Customs and Laws Drive Preventive Maintenance	Electronics TOW Components, TOW Test Training Equipment and Simulators (TOW) Dragon Components Dragon Test Simulators
29E	First Aid Navigate NBC Weapons Field Techniques Communications A/A (ID) Drive Preventive Maintenance	Electronics Radios Radio Maintenance Specialty/Test Equipment Troubleshooting Equipment Repair

(Continued)

Table 4.6 (Continued)

Functional Categories Comprising the Two Subscores for the EOT School Knowledge Tests

MOS	Basic	Technical
31C	First Aid Navigate NBC Weapons Field Techniques A/A (ID) Customs and Laws Drive Preventive Maintenance Vehicle Operation/Recovery	Generators Maintain TTY Electronic Equipment TTY Station and Net Operations Operate TTY Electronic Equipment Install Electronic Equipment
51B	First Aid Navigate NBC Weapons Field Techniques Communications A/A (ID) A/A (Engage) Customs and Laws	Construction Math Construction Tools and Materials Rigging Demolition Field Fortifications and Obstacles
54E	First Aid Navigate Weapons Field Techniques Communications A/A (ID) A/A (Engage) Customs and Laws Drive Preventive Maintenance Generators	NBC Chemical/Biological/Radiological Operations Decontamination Operations and Equipment
55B	First Aid Navigate NBC Weapons Field Techniques Communications	Ammunition Markings/Characteristics Ammunition Demolition Storage/Transport Symbols and Signs Ammunition Administration

(Continued)

Table 4.6 (Continued)

Functional Categories Comprising the Two Subscores for the EOT School Knowledge Tests

MOS	Basic	Technical
55B	A/A (ID) A/A (Engage) Customs and Laws Drive Preventive Maintenance	
63B	First Aid Navigate NBC Weapons Field Techniques Communications Customs and Laws Drive Preventive Maintenance	Vehicle Operation/Recovery Electrical System Brake/Steering/Suspension Systems Fuel/Cooling/Lubricating Systems
67N	First Aid Navigate NBC Weapons Field Techniques Communications A/A (ID) Customs and Laws	Aircraft Maintenance Administration Helicopter Maintenance Landing/Takeoff Signals Aircraft Refuel Aircraft Maintenance Tools/Equipment
71L	First Aid Navigate NBC Weapons Communications A/A (ID) Customs and Laws Drive	Forms/Files Management Supervision/Coordination Correspondence Classified Material
76Y	First Aid Navigate NBC Weapons Field Techniques Communications	Personnel/Organization Clothing/Individual Equipment Supply Administration Supply Storage Property Book Accounting Unit Supply Management

(Continued)

Table 4.6 (Continued)

Functional Categories Comprising the Two Subscores for the EOT School Knowledge Tests

MOS	Basic	Technical
76Y	A/A (ID) A/A (Engage) Customs and Laws Drive Preventive Maintenance	Unit Armory
88M	No Basic Categories	First Aid Navigate NBC Weapons Field Techniques Communications A/A (Engage) Customs and Laws Drive Preventive Maintenance Vehicle Operation/Recovery
91A	Navigate NBC Field Techniques Drive	First Aid Clinic/Ward Treatment and Care Clinic/Ward Housekeeping Clinic/Ward Management General Medical Knowledge
94B	First Aid Navigate NBC Weapons Field Techniques Communications Customs and Laws Preventive Maintenance	General Beverage Non-Meat Food Preparation Meat Preparation Food Service Food Service Administration
95B	First Aid Navigate NBC Weapons Field Techniques Communications Customs and Laws Drive Preventive Maintenance Vehicle Operation/Recovery	Conduct MP Procedures Patrol Duties Responding to Alarms

(Continued)

Table 4.6 (Continued)

Functional Categories Comprising the Two Subscores for the EOT School Knowledge Tests

MOS	Basic	Technical
96B	First Aid NBC Weapons Field Techniques Communications A/A (ID) Drive	Navigate Overlays Tactics/Tactical Intelligence Information Situation Maps Intelligence Reports Security Clearance/Administration Intelligence Briefings/Concepts Intelligence Communications

^aA/A = Antiaircraft/Antitank

END-OF-TRAINING RATINGS

This section describes preliminary analyses of the EOT Army-wide performance ratings. The purposes of these analyses were to (a) examine the distributional and psychometric properties of the individual EOT rating scales, and (b) combine the individual scales into higher level rating factor scores. These analyses include the computation of descriptive statistics (means, standard deviations, reliabilities), as well as the use of exploratory and confirmatory factor analysis techniques.

The EOT ratings consist of seven individual scales, all of which are modified versions of the Army-wide Behaviorally Anchored Rating Scales (BARS) developed to assess first-tour job performance (see Pulakos and Borman, 1986). These scales, as well as the defining concept provided with each one, are listed in Figure 4.1. Ratings were provided by peers (classmates) and supervisors (drill instructors). Each rating was on a 7-point scale.

The EOT ratings data set initially included data associated with 193,932 rater/ratee pairs across the 21 MOS. These ratings were distributed among 44,097 different ratees. The average number of peer raters per ratee was 3.54 (SD = 1.23), and the average number of supervisor raters per ratee was 0.86 (SD = 0.37). Table 4.7 shows these numbers for the total sample as well as separately by MOS.

Rating Scale 1: Technical Knowledge/Skill

- How effective is each soldier in acquiring job/soldiering knowledge and skill?

Rating Scale 2: Effort

- How effective is each soldier in displaying extra effort?

Rating Scale 3: Following Regulations and Orders

- How effective is each soldier in adhering to regulations, orders, and SOP and displaying respect for superiors?

Rating Scale 4: Military Appearance

- How effective is each soldier in maintaining proper military appearance?

Rating Scale 5: Physical Fitness

- How effective is each soldier in maintaining military standards of physical fitness?

Rating Scale 6: Self-Control

- How effective is each soldier in controlling own behavior related to aggressive acts?

Rating Scale 7: Leadership Potential

- On this last rating scale, evaluate each soldier on his or her potential effectiveness as a leader. At this point you are not necessarily to rate on the basis of present performance, but instead to indicate how well each soldier is likely to perform in leadership positions in his or her MOS.

Figure 4.1. End-of-Training Army-wide performance rating scales.

Table 4.7**Mean Peer and Supervisor EOT Ratings per Ratee, by MOS: Unedited Data**

MOS	Ratees	<u>Peer Ratings per Ratee</u>		<u>Supv Ratings per Ratee</u>	
		Mean	SD	Mean	SD
11B	10298	3.61	.74	.96	.19
12B	2011	3.14	1.37	.79	.41
13B	5288	3.94	.42	.96	.19
16S	691	3.96	.28	.94	.23
19E	481	3.82	.65	.97	.17
19K	1655	3.90	.42	.97	.17
27E	180	2.79	1.40	.44	.51
29E	306	3.68	.95	.83	.39
31C	1358	3.78	.72	.90	.31
51B	391	1.75	1.37	.41	.49
54E	805	4.72	.89	.98	.13
55B	690	3.44	1.19	.63	.53
63B	1474	2.18	1.62	.54	.51
67N	407	4.06	.65	1.00	.00
71L	1838	3.36	1.13	.84	.42
76Y	2281	2.75	1.76	.63	.56
88M	1933	2.27	1.60	.51	.51
91A	5334	3.89	.58	.98	.16
94B	2660	2.18	1.67	.52	.54
95B	3766	4.79	.67	.92	.34
96B	250	3.88	.72	.72	.45
Total	44097	3.54	1.23	.86	.37

Data Editing and Outlier Analyses

The EOT ratings analyses were conducted using three specially constructed samples, which will be referred to as the Preliminary Analysis sample, the Primary Analysis sample, and the Supervisor Ratings Reliability sample. Before these samples were drawn, however, the original EOT Ratings data set was edited (and data were deleted) according to three criteria: (a) excessive missing data for individual peer rater/ratee pairs; (b) mismatching MOS codes for peer rater/ratee pairs; and (c) rater/ratee pairs for which the rater (peer or supervisor) supplying the ratings was identified as an outlier. The results of these data edits are described below.

Excessive Missing Data. Before any outlier analyses were conducted, all peer rater/ratee pairs with four or more missing ratings (out of seven) were deleted from the original data set. This resulted in the elimination of 510 rater/ratee pairs (less than one-third of one percent of the total number of peer rater/ratee pairs). This criterion was not applied to supervisor rater/ratee pairs because the majority of ratees were evaluated by only one supervisor, so elimination of such supervisor ratings would have resulted in the deletion of all supervisory-level information for most of the ratees affected. However, such information was needed for the outlier analyses described below.

Mismatching MOS Codes. In addition to the peer rater/ratee pairs with excessive missing data, peer rater/ratee pairs with mismatching MOS codes were also deleted from the original data set. The purpose of this criterion was to ensure that all peer ratings included in the analyses were assigned by raters in the same MOS as the individuals being evaluated. A total of 521 rater/ratee pairs (again less than one-third of one percent of all peer rater/ratee pairs) were eliminated because the MOS of the rater was not the same as the MOS of the ratee.

Outlier Analyses: Peer Raters. Following the elimination of peer rater/ratee pairs according to the above criteria, three outlier indexes were constructed for each peer rater who provided ratings for three or more ratees. For two of these indexes, we first computed the average rating for each rater/ratee pair. We then took these average ratings and computed the mean and standard deviation within each rater. The mean provided an indication of the harshness/leniency of each rater, and the standard deviation provided an indication of the extent to which each rater differentiated among the ratees that he or she evaluated.

After the resulting distributions of scores on these two indexes were examined, the decision was made to flag all raters whose mean average rating across three or more ratees was equal to either one or seven. These raters were flagged because they failed to differentiate between both scale dimensions and ratees, and because they used only one or the other of the two extreme values on the seven scales.

A third outlier index, constructed from two subscores, was also developed for identifying potential peer rater outliers. The first subscore was based on the squared deviations between the ratings assigned by all other peer raters to the exact same ratees. (For example, if Peer Rater X assigned Ratee W a score of 3 on the first rating scale, but Peer Raters Y and Z each assigned that ratee a score of 5, then the squared deviation for Rater X from

the average of raters Y and Z on Scale 1 for Ratee W would be equal to $(5-3)^2$, or 4.) For each rater, the first subscore was computed as the average squared deviation across all seven scales and across all ratees evaluated by that rater. This subscore provided an indication of the degree of agreement between each peer rater and his or her peers with regard to applying the EOT rating scales. Similarly, the second subscore was the average of the squared deviations between the ratings assigned by a given rater and the average of the ratings assigned by supervisor raters to the same ratees. The third index was the sum of the two squared deviation subscores.

After examining the resulting distribution of scores for this index, we decided to flag all raters scoring four or more standard deviations above the mean. To score this high on this index, they would have had to deviate substantially from both the peer and the supervisor raters evaluating the same ratees.

A total of 270 peer raters (out of 41,553) were identified by at least one of the flags described above. These raters provided ratings associated with a total of 932 rater/ratee pairs. Each of these pairs was eliminated from the data set.

Outlier Analysis: Supervisor Ratets. Only the first two indexes described above (i.e., the mean and standard deviation of the average ratings assigned by each rater to all of the ratees that he or she evaluated) were used to identify potential supervisor rater outliers. Once again, these indexes were computed for only those raters who had provided ratings for at least three ratees. Only one supervisor rater (out of 1,448) was flagged according to scores on these indexes. That supervisor evaluated five ratees and assigned each of them a 7 on each of the seven rating scales. All five rater/ratee pairs were deleted from the data set.

The third index, the sum of averaged squared deviation scores, was not used to eliminate supervisor raters. Since very few ratees were evaluated by multiple supervisors, such a score would have to have been computed using deviations from average peer ratings only. However, since supervisor ratings may deviate from peer ratings for many valid reasons (e.g., different organizational perspective), such deviations in and of themselves did not seem sufficient evidence to warrant identifying the supervisor raters as outliers.

Edited Data Set. Following the edits described above, 191,964 rater/ratee pairs remained on the EOT ratings data set. These ratings were distributed among 44,059 ratees. The average number of peer raters per ratee was 3.50 (SD = 1.21), and the average number of supervisors per ratee was 0.86 (SD = 0.37). Table 4.8 shows these numbers for the total sample as well as separately by MOS.

EOT Ratings Analysis Samples

As previously indicated, the analyses described in this section were conducted using one or more of three samples drawn from the edited data set described above. The creation of these samples (the Preliminary Analysis sample, the Primary Analysis sample, and the Supervisor Ratings Reliability sample) is described below.

Table 4.8**Mean Peer and Supervisor EOT Ratings per Ratee, by MOS: Edited Data**

MOS	Ratees	<u>Peer Ratings per Ratee</u>		<u>Supv Ratings per Ratee</u>	
		Mean	SD	Mean	SD
11B	10298	3.59	.75	.96	.19
12B	2011	3.11	1.35	.79	.41
13B	5288	3.89	.47	.97	.19
16S	691	3.91	.36	.94	.23
19E	481	3.81	.66	.97	.17
19K	1655	3.89	.43	.97	.17
27E	180	2.74	1.31	.44	.51
29E	306	3.65	.95	.83	.39
31C	1358	3.76	.72	.90	.31
51B	387	1.76	1.37	.41	.49
54E	805	4.66	.94	.98	.13
55B	690	3.37	1.20	.63	.53
63B	1474	2.12	1.52	.54	.51
67N	407	4.05	.66	1.00	.00
71L	1831	3.30	1.07	.84	.42
76Y	2267	2.60	1.50	.64	.56
88M	1929	2.23	1.52	.51	.51
91A	5334	3.86	.58	.98	.16
94B	2652	2.08	1.49	.53	.54
95B	3766	4.77	.69	.92	.34
96B	250	3.77	.78	.72	.45
Total	44059	3.50	1.21	.86	.37

Preliminary Analysis Sample. The Preliminary Analysis sample was comprised of 100 ratees (randomly selected) per MOS, each of whom had received complete sets of EOT ratings from at least two different peer raters and one supervisor rater. (See Table 4.9 for a list of the total number of ratees meeting this requirement in each MOS.) Only ratees from MOS with at least 600 ratees meeting this requirement were included in this sample. These MOS are 11B, 12B, 13B, 19K, 31C, 54E, 63B, 71L, 76Y, 88M, 91A, 94B, and 95B. For ratees who had been rated by more than two peers and/or more than one supervisor, exactly two sets of peer ratings and one set of supervisor ratings were randomly selected for inclusion in the sample. Altogether, the Preliminary Analysis sample consisted of 1,400 ratees and 4,200 rater/ratee pairs.

This sample was primarily created to be used to explore different factor models potentially underlying the EOT rating scales. These models could then be subjected to confirmatory factor analysis techniques in conjunction with the Primary Analysis sample (described below).

Primary Analysis Sample. The Primary Analysis sample was comprised of all ratees in each MOS who had complete sets of EOT ratings from two peers and one supervisor rater and who were not randomly assigned to the Preliminary Analysis sample. The Primary sample was created for use in the majority of the present analyses, including the computation of descriptive statistics (means, standard deviations, reliabilities), and the conduct of confirmatory factor analyses on models based on results of exploratory analyses of data in the Preliminary Analysis sample. The Primary Analysis sample included 34,442 ratees and 103,326 rater/ratee pairs. Of 21 MOS, only three (27E, $n = 75$; 51B, $n = 120$; and 96B, $n = 176$) were represented by ratee subsamples smaller than 200.

Supervisor Ratings Reliability Sample. As indicated above, the EOT ratings data set includes a large number of ratees who were rated by more than one peer, but a rather modest sample of ratees who were likewise rated by more than one supervisor. (See Table 4.10 for a listing of the number of ratees in each MOS with at least two complete sets of supervisor ratings.) Consequently, the previously described samples did not exclude ratees with only one set of supervisor ratings. However, estimates of interrater reliability require the collection and examination of ratings obtained from multiple raters of the same ratees. Therefore, to study the reliability of EOT supervisor ratings, a third analysis sample was created. This sample was comprised of all 246 ratees for whom complete ratings were supplied by two or more supervisors. Again, for those ratees with more than two complete sets of supervisor ratings, two sets were randomly selected for inclusion in the sample.

Descriptive Statistics and Reliabilities

The descriptive statistics for the EOT ratings, based on the primary sample, are shown below. Reliability estimates are based on both the primary sample and the supervisor sample.

Table 4.9

Number of Ratees With a Minimum of Two Complete Sets of Peer EOT Ratings and One Complete Set of Supervisor EOT Ratings, by MOS

MOS	Ratees	MOS	Ratees
11B	9653	55B	391
12B	1478	63B	628
13B	5052	67N	404
16S	639	71L	1464
19E	457	76Y	1180
19K	1593	88M	828
27E	75	91A	5139
29E	250	94B	1025
31C	1199	95B	3330
51B	120	96B	176
54E	761	Total	35842

Table 4.10

Number of Ratees with a Minimum of Two Complete Sets of Supervisor EOT Ratings, by MOS

MOS	Ratees	MOS	Ratees
11B	1	55B	14
12B	4	63B	11
13B	9	67N	0
16S	0	71L	22
19E	0	76Y	45
19K	0	88M	10
27E	1	91A	14
29E	1	94B	48
31C	4	95B	62
51B	0	96B	0
54E	0	Total	246

Means and Standard Deviations

The results of these analyses for the total sample are reported in Table 4.11. Means and standard deviations were computed separately by rater type for each of the seven rating scales.

Table 4.11

Means of EOT Rating Scales: Primary Analysis Sample

Rating Scale ^a	Raters					
	Peer 1		Peer 2		Supervisor	
	Mean	SD	Mean	SD	Mean	SD
1	4.51	1.30	4.51	1.30	4.88	1.20
2	4.31	1.50	4.31	1.51	4.80	1.35
3	4.52	1.49	4.52	1.50	4.93	1.38
4	4.73	1.35	4.73	1.35	4.78	1.26
5	4.72	1.41	4.72	1.40	4.87	1.32
6	4.65	1.68	4.64	1.68	4.95	1.36
7	4.05	1.69	4.06	1.70	4.50	1.50
Average	4.50	1.49	4.50	1.49	4.82	1.34

Note: N = 34,442.

^aEach scale was rated from 1 to 7.

Note that results for peer ratings are reported twice, once under the column labeled "Peer 1" and once under the column labeled "Peer 2". These columns represent replications of peer rating results made possible by the inclusion of two sets of peer ratings for each ratee in the sample. For each ratee, one set of peer ratings was randomly assigned to Peer Group 1 and the other to Peer Group 2. Since the exact same set of ratees were evaluated by the raters in the two peer rater groups (as well as in the one supervisor rater group), differences between the results in the columns can be attributable to differences between raters. Furthermore, to the extent that the results for the two groups of peer raters are more similar to each other than they are to the results for the supervisor raters, there is evidence for differences between the ratings due to rater type.

The results in Table 4.11 suggest that peer and supervisor raters each made use of reasonable portions of the 7-point scales. (The standard deviations ranged from 1.30 to 1.70 for the peer ratings, and from 1.20 to 1.50 for the supervisor ratings.) Additionally, the means and standard deviations of the ratings for the two sets of peer raters were more similar to each other than to the means and standard deviations of the supervisor ratings. The mean peer ratings were lower than the mean supervisor ratings,

but their standard deviations were slightly higher. These differences are consistent across all seven rating scales, although they are less pronounced for rating scales 4 and 5 (Military Appearance and Physical Fitness).

Reliabilities

Reliabilities of the EOT peer ratings were examined using the Primary Analysis sample. Single-rater and two-rater reliabilities for the total sample are reported in Table 4.12 under the column labeled "Peer 1-Peer 2". The single-rater reliabilities were estimated using the intraclass ratings. Two-rater reliabilites were derived using the Spearman-Brown correction. The average single-rater reliability of the peer ratings across the seven scales was .32, and the average two-rater reliability was .48.

Although only 246 ratees on the EOT Ratings data set were rated by two or more supervisors, single-rater and two-rater reliabilities were estimated for these ratees, using the Supervisor Ratings Reliability sample. The results of these analyses are reported in Table 4.13. The average single-rater reliability of the supervisor ratings across the seven scales was only .14, and the average two-rater reliability only .25.

The results in Table 4.13 should be considered as no more than lower bound estimates of reliability for the EOT supervisor ratings. It is not clear why the 246 ratees in the Supervisor Ratings Reliability sample received multiple supervisor evaluations in the first place. The supervisor ratings were supposed to be collected from each ratee's drill instructor, so there should not have been more than one set of ratings per ratee. One possibility is that a small number of supervisors who were not drill instructors managed to provide ratings anyway. These supervisors may not have been as familiar with the ratees as the drill instructors were. Another possibility is that some of these ratees had "recycled" during the time span when data were being collected. These ratees may have performed poorly in training the first time around, and improved their performance the second time around. Whatever the explanation for the existence of these multiple supervisor ratings, the results associated with them should be interpreted with caution.

Alternative lower bound estimates of reliability for the EOT supervisor ratings can be found in Table 4.12. The columns labeled "Peer 1-Supervisor" and "Peer 2-Supervisor" contain intraclass correlations between supervisor ratings and ratings for each of the two groups of peer raters, respectively. Although these estimates are not what are usually considered estimates of reliability (in that the ratings being correlated were obtained from two distinct sources of raters), they do provide a lower estimate of the systematic variance contained in the supervisor (and peer) ratings. To the extent that the supervisors shared some type of rating "policy" which somehow differed from that used by peers, the estimates reported in Table 4.12 are underestimates of the actual reliabilities of the supervisor ratings. The average single-rater reliability across the seven rating scales (as indicated by both the "Peer 1-Supervisor" and "Peer 2-Supervisor" intraclass correlations) was .23, and the average two-rater reliability was .37.

Table 4.12**Single-Rater/Two-Rater Reliabilities of EOT Rating Scales: Primary Analysis Sample**

Rating Scale	Rater Pairs		
	Peer 1- Peer 2	Peer 1- Supervisor	Peer 2- Supervisor
1	.31/.47	.20/.34	.20/.34
2	.28/.43	.19/.32	.19/.32
3	.31/.47	.24/.38	.24/.38
4	.26/.41	.20/.34	.20/.34
5	.43/.60	.35/.51	.35/.51
6	.34/.51	.18/.30	.17/.30
7	.32/.49	.26/.41	.26/.41
Average	.32/.48	.23/.37	.23/.37

Note: N = 34,442.

Table 4.13**Single-Rater/Two-Rater Reliabilities of EOT Rating Scales: Supervisor Ratings Reliability Sample**

Rating Scale	Supervisor 1- Supervisor 2
1	-.01/-.01
2	.12/.22
3	.12/.22
4	.15/.26
5	.25/.39
6	.22/.36
7	.16/.28
Average	.14/.25

Note: N = 246.

Exploratory Factor Analyses

As previously indicated, factor analyses of the EOT ratings were conducted in two steps. The first step consisted of exploratory analyses using the Preliminary Analysis sample. The purpose of this step was to examine potential factor models underlying the seven EOT ratings scales. Least squares principal factor analyses based on the total sample were conducted separately for the three sets of ratings (Peer 1, Peer 2, and Supervisor). The eigenvalues and proportion of common variance explained by each of the first seven factors (unrotated) are reported in Table 4.14. The results show that the proportion of common variance accounted for by the first factor exceeded 1.00 for all three sets of ratings, suggesting that one factor may be sufficient to account for the common variance associated with the ratings assigned by both the peer and supervisor raters.

Table 4.15 reports the factor loadings for the one-factor solution computed separately for the ratings associated with each of the two sets of peer raters as well as the supervisor ratings. Note that the loadings for the supervisor ratings are consistently greater than the loadings for both sets of peer ratings. This suggests that the level of correlations among the supervisor ratings was greater than the level of correlations among the peer ratings.

Confirmatory Factor Analyses

The results of the exploratory analyses provide strong indication of a one factor model for both peer and supervisor ratings. However, prior research conducted with the 10 Army-wide Behaviorally Anchored Rating Scales (from which the first six EOT ratings were selected and modified) found multiple factors to underlie these scales. More specifically, in analyses of data collected during the measurement of first-tour performance in the Concurrent Validation phase of Project A, three factors (labeled Effort and Leadership, Maintaining Personal Discipline, and Physical Fitness and Military Bearing) were found to underlie the 10 Army-wide ratings when they were used to evaluate on-the-job performance (Pulakos & Borman, 1986). As a result of these analyses, the two highest loading scales on each of the three factors were administered in the present research effort. The correspondence between the first six EOT rating scales and the three factors is as follows:

- Effort and Leadership - Scales 1 and 2
- Personal Discipline - Scales 3 and 6
- Physical Fitness and Military Bearing - Scales 4 and 5

(The seventh EOT rating scale, Leadership Potential, corresponds to an 11th Army-wide rating scale, labeled NCO Potential, which was also administered during the CV data collection. That scale, however, was not included in the Project A factor analyses.)

In the present analyses, a decision was made to compare the one-factor model suggested by the results of the exploratory factor analyses with a model corresponding to the results found in Project A. This latter model was specified as consisting of four oblique factors, including three two-scale factors (as described above) and a single-scale factor of Leadership Potential.

Table 4.14**Exploratory Factor Analysis of EOT Rating Scales: Proportion of Common Variance Accounted for by First Seven Unrotated Factors^a**

	Eigenvalues	Proportion of Common Variance Explained
<u>Peer 1</u>		
Factor 1	3.26	1.07
Factor 2	.23	.08
Factor 3	.03	.01
Factor 4	-.06	-.02
Factor 5	-.10	-.03
Factor 6	-.14	-.05
Factor 7	-.18	-.06
<u>Peer 2</u>		
Factor 1	3.13	1.10
Factor 2	.21	.07
Factor 3	-.01	.00
Factor 4	-.07	-.02
Factor 5	-.11	-.04
Factor 6	-.13	-.05
Factor 7	-.17	-.06
<u>Supervisor</u>		
Factor 1	4.29	1.05
Factor 2	.12	.03
Factor 3	.01	.00
Factor 4	-.04	-.01
Factor 5	-.08	-.02
Factor 6	-.11	-.03
Factor 7	-.12	-.03

Note: Based on Preliminary Analysis sample, N = 1,400.

^aComputed separately for Peer 1, Peer 2, and Supervisor ratings.

Table 4.15**Exploratory Factor Analysis of EOT Rating Scales: Factor Loadings for One-Factor Solutions^a**

Rating Scale	Peer 1 Factor 1	Peer 2 Factor 1	Supervisor Factor 1
1	.75	.72	.77
2	.73	.76	.85
3	.72	.70	.83
4	.66	.64	.78
5	.54	.53	.61
6	.59	.54	.78
7	.76	.74	.84

Note: Based on Preliminary Analysis sample, N = 1,400.

^aComputed separately for Peer 1, Peer 2, and Supervisor ratings.

To determine which of the two models appeared most plausible for the data in the Primary Analysis sample, each was subjected to a series of confirmatory factor analyses using LISREL. Three matrixes, referred to in LISREL as Lambda Y (LY), Psi (PS), and Theta Epsilon (TE), were estimated for each model. These matrixes contain, respectively, estimates of the factor loadings of each scale onto its assigned factor, estimates of the correlations among the factors, and estimates of the unique variance associated with each scale.

As previously indicated, LISREL provides several statistical measures of the fit of a model to a set of data. Among these measures are the chi-square value for the model, the root mean square residual (RMS), the goodness-of-fit index (GFI), and the adjusted goodness-of-fit index (AGFI). Chi-square values can be used to statistically compare the fit of models that are nested. More specifically, when two models are nested, the difference between their respective chi-square values is itself distributed according to chi-square (with degrees of freedom equal to the difference between the degrees of freedom associated with each of the two models). If the difference between the two chi-square values is significant, then this indicates that the less restrictive model (i.e., the model with the fewer degrees of freedom) provides a significantly better fit to the data.

Peer 1 Ratings

Table 4.16 reports the results of the comparison between the one- and four-factor models for the ratings associated with the raters in the Peer 1 group (using the total Primary Analysis sample). The four-factor model provides a better fit to the peer ratings than does the more restrictive one-factor model. In addition to the fact that the difference between the chi-

square values for the two models was highly significant, the results also show that the root mean square residual was reduced from .04 for the one-factor model to .02 for the four-factor model. Likewise, the adjusted goodness-of-fit index (which is adjusted for degrees of freedom in the model) increased from .95 to .97. Finally, the ratio of chi-square to degrees of freedom for the one-factor model was almost twice as great as it was for the four-factor model.

Table 4.16

Confirmatory Factor Analysis: Comparison of Fit of One- and Four-Factor Models, Based on Peer 1 EOT Ratings

Model	df	GFI	AGFI	RMS	χ^2
1 Factor	14	.97	.95	.04	3118.7
4 Factors	10	.99	.97	.02	1156.4
Difference	4	-	-	-	1962.4

Note: Based on Primary Analysis sample, N = 34,442.

Table 4.17 reports the matrixes of factor loadings and factor correlations which were estimated for the four-factor model as applied to the Peer 1 ratings. (Asterisks denote values that were constrained by the model.) The correlations among the factors ranged between .68 and .88. These figures represent correlations between latent variables and, as such, have already been corrected for attenuation due to measurement error.

Confirmatory factor analyses for the one- and four-factor models were also conducted for the Peer 1 ratings separately for each MOS in the Primary Analysis sample. The results of these analyses are reported in Tables 4.18 and 4.19 respectively. These results suggest that the improved fit associated with the four-factor model is consistent across MOS.

Stability of Peer Rating Factor Structure

Because ratees in the Primary Analysis sample were each assigned ratings from two different peers (i.e., Peer 1 and 2), it was possible to examine the stability of the peer rating factor structure identified above. The section of Table 4.20 labeled "Separately" reports fit indexes for the four-factor model estimated separately (using the total sample) for the Peer 1 and Peer 2 ratings. Note that the two sets of results are very similar. In particular, the root mean square residual for both sets of ratings was only .02.

Table 4.17

Confirmatory Factor Analysis: Pattern Matrix and Factor Correlation Matrix for Four-Factor Model, Based on Peer 1 EOT Ratings

Rating Scale	Factor ^a			
	Effort/ Leadership (ELS)	Personal Discipline (MPD)	Fitness/ Bearing (PFB)	Leadership Potential (LEAD)
<u>Pattern Matrix</u>				
1	.77	.00*	.00*	.00*
2	.79	.00*	.00*	.00*
3	.00*	.81	.00*	.00*
4	.00*	.00*	.73	.00*
5	.00*	.00*	.62	.00*
6	.00*	.63	.00*	.00*
7	.00*	.00*	.00*	1.00*
<u>Factor Correlation Matrix</u>				
ELS	1.00*			
MPD	.88	1.00*		
PFB	.84	.74	1.00*	
LEAD	.77	.68	.72	1.00*

Note: Based on Primary Analysis sample, N = 34,442.

^a * indicates constrained by model.

Table 4.18

Confirmatory Factor Analysis: Fit of One-Factor Model, Based on Peer 1 EOT Ratings, by MOS

MOS	N	df	GFI	AGFI	RMS	χ^2
11B	9553	14	.97	.95	.04	917.5
12B	1378	14	.96	.93	.05	178.8
13B	4952	14	.97	.94	.05	514.5
16S	539	14	.97	.93	.05	65.3
19E	457	14	.97	.93	.05	52.7
19K	1493	14	.97	.94	.05	148.4
27E	75	14	.91	.82	.10	23.2
29E	250	14	.96	.92	.04	34.7
31C	1099	14	.98	.96	.03	83.9
51B	102	14	.92	.83	.08	35.2
54E	661	14	.97	.94	.04	72.9
55B	391	14	.96	.93	.04	51.1
63B	528	14	.95	.89	.07	99.9
67N	404	14	.98	.96	.04	31.6
71L	1364	14	.97	.95	.04	132.3
76Y	1080	14	.97	.95	.04	103.6
88M	728	14	.96	.93	.06	92.2
91A	5039	14	.98	.96	.04	396.2
94B	925	14	.96	.93	.05	119.1
95B	3230	14	.97	.95	.04	304.4
96B	176	14	.94	.88	.05	35.5
Mean	-	-	.96	.92	.05	-
Total	34,442	294	-	-	-	3492.9

Note: Based on Primary Analysis sample.

Table 4.19

Confirmatory Factor Analysis: Fit of Four-Factor Model, Based on Peer 1 EOT Ratings, by MOS

MOS	N	df	GFI	AGFI	RMS	χ^2
11B	9553	10	.99	.96	.02	433.9
12B	1378	10	.99	.96	.03	71.4
13B	4952	10	.99	.97	.02	183.5
16S	539	10	.98	.95	.03	35.7
19E	457	10	.99	.96	.03	24.0
19K	1493	10	.99	.96	.03	75.0
27E	75	10	.96	.89	.06	9.8
29E	250	10	.98	.94	.03	17.5
31C	1099	10	1.00	.99	.01	18.6
51B	102	10	.95	.85	.06	23.0
54E	661	10	.99	.96	.03	35.3
55B	391	10	.99	.97	.02	15.0
63B	528	10	.98	.95	.03	32.1
67N	404	10	.99	.96	.03	18.2
71L	1364	10	1.00	.99	.01	21.4
76Y	1080	10	.99	.98	.02	31.9
88M	728	10	.99	.98	.02	19.3
91A	5039	10	.99	.98	.02	113.9
94B	925	10	.99	.96	.02	41.7
95B	3230	10	.99	.96	.02	143.8
96B	176	10	.96	.90	.04	22.3
Mean	-	-	.98	.96	.03	-
Total	34,442	210	-	-	-	1387.4

Note: Based on Primary Analysis sample.

To determine whether the parameter estimates (i.e., elements in the LY, PS, and TE matrixes) underlying the two sets of results were significantly different from one another, the four-factor model was also estimated simultaneously for the Peer 1 and Peer 2 ratings using the LISREL multigroup option. For this analysis, the parameters contained in each of the three matrixes were constrained to be invariant (i.e., equal) across the two sets of peer ratings. The results of this analysis are reported in the section labeled "Simultaneously" in Table 4.20.

The results in Table 4.20 indicate that the additional constraints specified in the latter analysis did not significantly reduce the fit of the four-factor model for the two sets of peer ratings. Note that the increase in chi-square (from the sum of the two chi-squares associated with the models estimated separately to the single chi-square associated with the models estimated simultaneously) is only 17.48. This difference corresponds almost exactly to the difference in the degrees of freedom associated with the two analyses (difference in degrees of freedom = 18). Also, the root mean square residual for each set of ratings was .02, regardless of whether the models were estimated separately or simultaneously.

Table 4.20

Confirmatory Factor Analysis: Comparison of Separately and Simultaneously Estimated Solutions of Four-Factor Model for Peer 1 and Peer 2 EOT Ratings

Rater	df	GFI	AGFI	RMS	χ^2
<u>Separately</u>					
Peer 1	10	.99	.97	.02	1156.4
Peer 2	10	.99	.98	.02	1044.1
Total	20	-	-	-	2,200.5
<u>Simultaneously</u>					
Peer 1	-	.99	-	.02	-
Peer 2	-	.99	-	.02	-
Total	38	-	-	-	2,218.0
$\chi^2_{(38)} - \chi^2_{(20)} = 2,218.01 - 2,200.50$ $\chi^2_{(18)} = 17.48$					

Note: Based on Primary Analysis sample, N = 34,442.

Supervisor Ratings

The complete set of confirmatory analyses reported in Tables 4.16-4.19 for the first set of peer ratings were conducted for the supervisor ratings as well. Table 4.21 reports the results of the total sample comparison between the one- and four-factor models for the supervisor ratings. Once again, the chi-square values, the root mean square residuals, and the adjusted goodness-of-fit indexes indicate that the four-factor model provides a better fit to the data than does the one-factor model.

Table 4.21

Confirmatory Factor Analysis: Comparison of Fit of One- and Four-Factor Models, Based on Supervisor EOT Ratings

Model	df	GFI	AGFI	RMS	χ^2
1 Factor	14	.96	.93	.03	4375.0
4 Factors	10	.99	.96	.02	1738.1
Difference	4	-	-	-	2636.9

Note: Based on Primary Analysis sample, N = 34,442.

Table 4.22 reports the pattern and factor correlation matrixes for the four-factor model as applied to the supervisor ratings. Note that the non-constrained values in both of these matrixes are somewhat higher than the corresponding values reported in Table 4.17. For example, whereas the average correlation among the factors was only .77 for the peer ratings, it was .85 for the supervisor ratings. These results suggest that the ratings provided by the supervisor raters were less differentiated than those provided by the peer raters.

Finally, confirmatory factor analyses for the one- and four-factor models were conducted separately by MOS for the supervisor ratings. The results of these analyses are reported in Tables 4.23 and 4.24. Once more, the results suggest that the four-factor model provides a better fit to the data than does the one-factor model, regardless of MOS.

Table 4.22

Confirmatory Factor Analysis: Pattern Matrix and Factor Correlation Matrix for Four-Factor Model, Based on Supervisor EOT Ratings

Rating Scale	Factor ^a			
	ELS	MPD	PFB	LEAD
<u>Pattern Matrix</u>				
1	.81	.00*	.00*	.00*
2	.89	.00*	.00*	.00*
3	.00*	.88	.00*	.00*
4	.00*	.00*	.85	.00*
5	.00*	.00*	.71	.00*
6	.00*	.80	.00*	.00*
7	.00*	.00*	.00*	1.00*
<u>Factor Correlation Matrix</u>				
ELS	1.00*			
MPD	.91	1.00*		
PFB	.90	.87	1.00*	
LEAD	.82	.79	.80	1.00*

Note: Based on Primary Analysis sample, N = 34,442.

^a * indicates constrained by model.

Table 4.23**Confirmatory Factor Analysis: Fit of One-Factor Model, Based on Supervisor EOT Ratings, by MOS**

MOS	N	df	GFI	AGFI	RMS	χ^2
11B	9553	14	.96	.92	.04	1263.3
12B	1378	14	.96	.93	.05	172.0
13B	4952	14	.96	.92	.04	722.3
16S	539	14	.95	.91	.05	87.4
19E	457	14	.94	.89	.05	91.4
19K	1493	14	.95	.91	.04	243.5
27E	75	14	.85	.70	.07	39.0
29E	250	14	.96	.92	.04	36.4
31C	1099	14	.95	.90	.04	197.7
51B	102	14	.95	.89	.06	22.5
54E	661	14	.95	.90	.04	110.9
55B	391	14	.98	.96	.02	30.7
63B	528	14	.96	.93	.03	60.6
67N	404	14	.92	.84	.06	109.9
71L	1364	14	.96	.93	.04	179.0
76Y	1080	14	.96	.92	.05	156.7
88M	728	14	.96	.92	.04	106.0
91A	5039	14	.96	.93	.03	631.9
94B	925	14	.95	.91	.04	153.7
95B	3230	14	.96	.92	.04	475.0
96B	176	14	.95	.90	.04	30.5
Mean	-	-	.95	.90	.04	-
Total	34,442	294	-	-	-	4929.5

Note: Based on Primary Analysis sample.

Table 4.24

Confirmatory Factor Analysis: Fit of Four-Factor Model, Based on Supervisor EOT Ratings, by MOS

MOS	N	df	GFI	AGFI	RMS	χ^2
11B	9553	10	.98	.95	.02	611.3
12B	1378	10	.99	.98	.02	37.5
13B	4952	10	.98	.95	.02	299.0
16S	539	10	.98	.94	.02	41.6
19E	457	10	.98	.94	.02	32.8
19K	1493	10	.98	.95	.02	95.1
27E	75	10	.91	.74	.06	24.3
29E	250	10	.98	.93	.02	21.0
31C	1099	10	.97	.91	.03	120.9
51B	102	10	.99	.97	.02	3.9
54E	661	10	.99	.96	.02	33.9
55B	391	10	.99	.97	.01	15.3
63B	528	10	.98	.94	.02	38.8
67N	404	10	.97	.90	.04	50.0
71L	1364	10	.98	.95	.02	78.7
76Y	1080	10	.99	.96	.02	54.1
88M	728	10	.99	.97	.02	26.1
91A	5039	10	.99	.96	.02	214.7
94B	925	10	.98	.94	.02	72.1
95B	3230	10	.98	.94	.02	241.5
96B	176	10	.98	.94	.03	12.8
Mean	-	-	.98	.94	.02	-
Total	34,442	210	-	-	-	2152.2

Note: Based on Primary Analysis sample.

Comparability of Peer and Supervisor Rating Factor Structure

In addition to permitting comparison of the factor structures underlying the two sets of peer ratings (as reported above), the Primary Analysis sample also provided the opportunity to compare the factor structure of the peer ratings with that of the supervisor ratings. Reported in Table 4.25 in the section labeled "Separately" are fit indexes for the four-factor model estimated separately for the Peer 1 and supervisor ratings. Although the chi-square value associated with the supervisor ratings is somewhat higher than that associated with the peer ratings, the root mean square residuals are equally low (.02) for both.

Table 4.25

Confirmatory Factor Analysis: Comparison of Separately and Simultaneously Estimated Solutions of Four-Factor Model for Peer 1 and Supervisor Ratings

Rater	df	GFI	AGFI	RMS	χ^2
<u>Separately</u>					
Peer 1	10	.99	.97	.02	1156.4
Supervisor	10	.99	.96	.02	1738.1
Total	20	-	-	-	2,894.5
<u>Simultaneously (All parameters invariant)</u>					
Peer 1	-	.94	-	.09	-
Supervisor	-	.99	-	.07	-
Total	38	-	-	-	11,219.3

$$\chi^2_{(38)} - \chi^2_{(20)} = 11,219.3 - 2,894.5$$

$$\chi^2_{(18)} = 8,324.8$$

Simultaneously (Only pattern matrix invariant)

Peer 1	-	.99	-	.03	-
Supervisor	-	.99	-	.02	-
Total	22	-	-	-	3,176.3

$$\chi^2_{(22)} - \chi^2_{(20)} = 3,176.3 - 2,894.5$$

$$\chi^2_{(2)} = 281.9$$

Note: Based on Primary Analysis sample, N = 34,442.

To determine whether the parameter estimates underlying the two sets of results were significantly different from one another, the four-factor model was estimated simultaneously for the Peer 1 and supervisor ratings. Once again, using LISREL's multigroup option, the parameters contained in the LY, PS, and TE matrixes were constrained to be equal across the two sets of ratings. The results of this analysis are reported in Table 4.25 in the section labeled "Simultaneously (All parameters invariant)". These results indicate that the equality constraints significantly reduced the fit of the four-factor model for the two sets of ratings. Specifically, the difference between the chi-square value associated with the two models estimated simultaneously and the sum of the chi-square values for the two models estimated separately was 8,324.8 (difference in degrees of freedom = 18). Also, the root mean square residuals associated with the peer and supervisor results increased to .09 and .07, respectively.

The section of Table 4.25 labeled "Simultaneously (Only pattern matrix invariant)" reports the results of a further attempt to simultaneously estimate the four-factor model for the Peer 1 and supervisor ratings. In this analysis, only the factor loading matrix (LY) was constrained to be equal across the two sets of ratings (i.e., the correlations among the factors, as well as the uniqueness of the scales, were allowed to vary between the peer and supervisor ratings). Although the fit associated with this analysis was still significantly worse than that associated with the four-factor models estimated separately (difference in chi-square = 281.88; difference in chi-square degrees of freedom = 2), the absolute reduction in fit was small. In particular, the root mean square residuals associated with the peer and supervisor ratings were .03 and .02, respectively. These results suggest that most of the differences between the factor structures underlying the peer and supervisor ratings are associated with the greater level of covariation among the supervisor ratings.

Creation of EOT Rating Factor Scores

The preceding analyses support the following decisions regarding the creation of EOT rating factor scores. First, the seven rating scales will be summarized into four separate rating factors. The confirmatory factor analyses consistently demonstrated (across MOS, across raters) the superiority of the oblique four-factor solution over the one-factor solution suggested by the exploratory analyses. The factor scores will be created by unit weighting and averaging the scales associated with each factor as specified in the previous analyses. Specifically, Scales 1 and 2 will be averaged to form the first factor which (to avoid confusion with the seventh rating scale) is relabeled here as Effort and Technical Skill (ETS). Similarly, Scales 3 and 6 will be averaged to create the second rating factor, Maintaining Personal Discipline (MPD), as will Scales 4 and 5 to create the third rating factor, Physical Fitness and Military Bearing (PFB). The fourth rating factor, labeled Leadership Potential (LEAD), will consist solely of Scale 7.

A second decision concerns the separate treatment of peer and supervisor ratings. Separate rating factor scores will be created for ratings obtained by peer and supervisor raters, respectively. This decision is based on the hypothesis that there may be differences in the psychological constructs underlying ratings obtained from the two sources. This raises the possibility that the different ratings may represent different aspects of performance and,

therefore, may be predicted by different predictor constructs. Several findings reported above suggest that, in fact, there may be differences between the two sets of ratings. First, the means of the supervisor ratings are higher. Second, the two sets of peer ratings correlated higher with each other than either of them did with the supervisor ratings. Finally, the correlations among the supervisor rating factors were higher than the correlations among the peer rating factors.

Distributional properties of the EOT rating factor scores were examined using the Primary Analysis sample. The results of these analyses for the total sample are reported in Table 4.26. Means and standard deviations were computed separately for both sets of peer ratings and the one set of supervisor ratings. The means of the supervisor rating factor scores were higher than the means of the peer rating factor scores, but the standard deviations of the supervisor rating factor scores were slightly lower than the standard deviation of the peer rating factor scores.

Table 4.26

Means of EOT Rating Factor Scores: Primary Analysis Sample

Rating Scale ^a	Raters					
	Peer 1		Peer 2		Supervisor	
	Mean	SD	Mean	SD	Mean	SD
ETS	4.41	1.25	4.41	1.26	4.84	1.18
MPD	4.58	1.38	4.58	1.38	4.94	1.27
PFB	4.72	1.17	4.72	1.17	4.83	1.15
LEAD	4.05	1.69	4.06	1.70	4.50	1.50

Note: N = 34,442.

Reliabilities of the peer rating factor scores (and pseudo-reliabilities of the supervisor rating factor scores) were also examined using the Primary Analysis sample. Estimates of single-rater and two-rater reliabilities of the peer ratings are reported in the first column of Table 4.27. The single-rater reliabilities ranged from .32 to .37, and the two-rater reliabilities ranged from .49 to .54.

The lower bound estimates of reliability for the supervisor rating factor scores (based on the intraclass correlations between the supervisor rating factor scores and the two sets of peer rating factor scores, respectively) are reported in the last two columns of Table 4.27. The single-rater reliabilities ranged from .23 to .30, and the two-rater reliabilities ranged from .37 to .46.

Table 4.27

Single-Rater/Two-Rater Reliabilities of EOT Rating Factor Scores: Primary Analysis Sample

Rating Scale	Rater Pairs		
	Peer 1- Peer 2	Peer 1- Supervisor	Peer 2- Supervisor
ETS	.34/.51	.23/.37	.23/.37
MPD	.37/.54	.24/.39	.24/.39
PFB	.37/.54	.30/.46	.30/.46
LEAD	.32/.49	.26/.41	.26/.41

Note: N = 34,442.

RELATIONSHIPS BETWEEN THE EOT FACTOR SCORES

The six EOT factor scores constitute the basic criterion scores for training performance. They correspond directly to the performance components identified in the CVI performance modeling analysis. To examine the relationships among the six EOT factor scores, correlations were obtained between the two EOT SK factor scores and the four rating scores, calculated separately for peers and supervisors, for the nine Batch A MOS (i.e., 11B, 13B/S and B/T, 19E, 31C, 63B, 71L, 88M, 91A, and 95B).

The correlations were obtained by MOS and then averaged, using a z -transformation, across MOS having the same number of EOT scores. Thus, the correlational analyses were run separately for two groups of Batch A MOS: (a) MOS 11B and 88M, having one SK factor score (i.e., Technical) and four rating scores, and (b) the seven other MOS that have scores on all six EOT factors. The correlations for these two groups between the SK factor scores and the ratings factor scores for peer raters are given in Tables 4.28 and 4.29. The same correlations for supervisor raters are given in Tables 4.30 and 4.31. These tables indicate that the correlations exhibit nearly identical patterns for both rater types, although they are higher for supervisor ratings than for peer ratings (almost certainly due to the higher reliability of the supervisor ratings).

Table 4.28

Average Correlations of EOT SK Test Scores With EOT Rating Scores Across Seven Batch A MOS^a Peer Raters

	BASIC	TECH	ETS	MPD	PFB	LEAD
BASIC	1.00					
TECH	.59	1.00				
ETS	.13	.18	1.00			
MPD	.10	.14	.63	1.00		
PFB	.02	.03	.57	.46	1.00	
LEAD	.09	.11	.65	.56	.57	1.00

^aThe seven Batch A MOS having two SK factor scores.

Table 4.29

Average Correlations of EOT SK Test Scores With EOT Rating Scores Across Two Batch A MOS^a Peer Raters

	BASIC	ETS	MPD	PFB	LEAD
BASIC	1.00				
ETS	.20	1.00			
MPD	.17	.60	1.00		
PFB	.07	.59	.45	1.00	
LEAD	.16	.67	.55	.56	1.00

^aMOS 11B and 88M, which have only one SK factor score.

Table 4.30

Average Correlations of EOT SK Test Scores With EOT Rating Scores Across Seven Batch A MOS^a Supervisor Raters

	BASIC	TECH	ETS	MPD	PFB	LEAD
BASIC	1.00					
TECH	.59	1.00				
ETS	.12	.14	1.00			
MPD	.11	.13	.77	1.00		
PFB	.06	.06	.73	.69	1.00	
LEAD	.09	.12	.74	.74	.71	1.00

^aThe seven Batch A MOS having two SK factor scores.

Table 4.31

Average Correlations of EOT SK Test Scores With EOT Rating Scores Across Two Batch A MOS^a Supervisor Raters

	BASIC	ETS	MPD	PFB	LEAD
BASIC	1.00				
ETS	.16	1.00			
MPD	.15	.73	.00		
PFB	.08	.72	.65	1.00	
LEAD	.16	.74	.68	.70	1.00

^aMOS 11B and 88M, which have only one SK factor score.

Chapter 5

Development of Scores for Second-Tour Performance Measures

This chapter will describe how the data from the Concurrent Validation II data collection were used to develop the basic scores for the second-tour performance measures. As background for the score development analyses, the general features of the second-tour performance measures and the characteristics of the CVII sample are briefly described below. Subsequent sections will deal with score development for each of the measures in turn. More detailed discussions of the development steps for each measure can be found in Campbell (1988) and Campbell and Zook (1990).

DEVELOPMENT OF THE SECOND-TOUR MEASURES

As described previously (Campbell, 1988), considerable job analysis information on which to base second-tour performance measurement was available. For each Batch A MOS, 30 technical (MOS-specific and common) tasks and 15 supervisory tasks were selected to represent the task clusters identified in the job analyses. The 45 tasks were then rank ordered in terms of their overall importance to the MOS. Critical incident analyses yielded a portrayal of each MOS in terms of its critical components in both technical performance and leadership. A series of job analysis interviews yielded an estimate of the relative importance and time spent for technical vs. supervisory activities for each MOS. Cluster analyses were used to further explore the specific dimensions of supervisory/leadership performance.

This analysis of second-tour jobs showed considerable overlap in job content between first tour and second tour, except that in the second tour the core technical tasks become more complex and significant components of leadership and supervision are introduced. Consequently, we modified a number of first-tour measurement methods for second-tour use, and we added several new measures of supervision and leadership.

Modifications of First-Tour Measures for Second-Tour Use

To accommodate the new supervisory measures, assessment of technical task knowledge and performance (i.e., hands-on and job knowledge tests) was allotted less time than in the first-tour performance assessment. Reducing assessment time was judged to be better than eliminating either measurement strategy because (a) highly reliable job knowledge tests can be written for almost any task, and (b) the hands-on tests were designed to have a high degree of content validity. For the job knowledge tests, testing time was reduced by using fewer items for each task. This strategy is not feasible with hands-on tests because the scorable steps within task tests are too interdependent to be selectively eliminated. Consequently, we tested fewer tasks in a hands-on mode relative to the number of tasks used to assess first-tour soldiers.

In addition to the hands-on and job knowledge tests, several types of rating scales and personnel records were used as methods of assessing performance for NCOs, as they were for first-term soldiers. The changes made

to the first-tour procedures for each type of measure are briefly summarized below, and discussed in more detail later in the chapter.

Rating Scales. The second-tour Army-wide and MOS-specific performance rating scales were developed using the first-tour scales as a starting point. Information generated through the second-tour job analysis was used to revise these instruments to make them suitable for second-tour soldiers. For example, in the Army-wide scales, the "NCO potential" scale was replaced with a "senior NCO potential" scale.

In addition, a set of scales was constructed to tap supervisory performance dimensions that were identified in the second-tour job analyses. A list of the areas covered by these scales and an example of one scale are provided in Figure 5.1.

The Army-wide, MOS-specific, and supervisory performance rating scales were administered during second-tour field testing. No changes to the scales were made as a result of analysis of those data.

A panel of subject matter experts indicated that the Combat Performance Prediction rating scales as revised for first-tour soldiers would also be applicable for second-tour soldiers. All of the rating scales intended for use with second-tour soldiers were administered during the field tests.

Hands-On and Job Knowledge Tests. By doctrine¹ Skill Level 2 soldiers are also responsible for Skill Level 1 tasks. Consequently, the technical tasks selected for testing first- and second-tour soldiers overlapped to a substantial degree. Development of new job knowledge and hands-on tests for the non-overlapping tasks was modeled after the procedures used for the first-tour tests. The tests were submitted to pilot testing and field testing before being finalized for administration to the second-tour sample. With respect to the job knowledge tests, item analyses on the field test data were used to identify items needing revision and to reduce the number of items so that the tests could be administered in one hour. Similarly, for the hands-on tests, field test results were used to identify needed revisions to the instructions and scorable steps of the hands-on tests. Also, the field test administration provided the information for determining which hands-on tests were to be administered and which were to be dropped.

Personnel File Form II. Personnel File Form II was developed by reviewing the contents of the Personnel File Form I with officers and NCOs who served as SME representatives for the Army's Military Personnel Center. In addition to the information on the first-tour version, the second-tour form elicits information from three categories: Education, Promotion Boards, and Reenlistment waivers. Army regulations were reviewed to identify information available on the Promotion Board Worksheet, and officers and NCOs who served on promotion boards were interviewed to provide more information about the NCO promotion process to E-5 and above. A draft version of the second-tour Personnel File Form was administered during the second-tour field tests. Only minor changes were made to the form as a result of field test data analyses.

¹Army Regulation 611-201, Enlisted Career Management Fields and Military Occupational Specialties.

Scale Areas

- 0 ACTING AS A ROLE MODEL
- 0 COMMUNICATION
- 0 PERSONAL COUNSELING
- 0 MONITORING SUBORDINATE PERFORMANCE
- 0 ORGANIZING MISSIONS/OPERATIONS
- 0 PERSONNEL ADMINISTRATION
- 0 PERFORMANCE COUNSELING/CORRECTING

ACTING AS A ROLE MODEL FOR SUBORDINATES

Motivates subordinates to perform effectively through personal example, including demonstrating high standards of military appearance, bearing, and courtesy; is a model supervisor for subordinates to look up to by demonstrating exemplary behavior as a soldier.

Falls below standards and expectations for performance in the category "Acting as a Model" compared to soldiers at same experience level.		Meets standards and expectations for performance in the category "Acting as a Model" compared to soldiers at same experience level.			Exceeds standards and expectations for performance in the category "Acting as a Model" compared to soldiers at same experience level.	
(1)	(2)	(3)	(4)	(5)	(6)	(7)

Figure 5.1. Example of supervisory/leadership performance ratings.

New Measurement Methods for Second-Tour Performance

Based on a review of the literature and a careful consideration of the feasibility of additional measurement methods, two new methods were developed for assessing second-tour NCO job performance: a situational judgment test and a series of supervisory simulation, or role-play, exercises. The simulation exercises were intended to assess the one-on-one interpersonal skills required for counseling and training subordinates, whereas the Situational Judgment Test (SJT) was intended to cover as broad a range of important supervisory skills as possible within the constraints of a paper-and-pencil format.

Situational Judgment Test (SJT)

The purpose of the SJT is to evaluate the effectiveness of judgments about what one should do in typical supervisory problem situations. A critical incident methodology was used to delineate situations to be included in the SJT. The SMEs who generated situations and response options were provided with the taxonomy of supervisory/leadership behaviors developed in the second-tour job descriptions.

Response options were formulated through a combination of input from pilot test SMEs and E-5 incumbents from field tests. SMEs wrote short answers (1-3 sentences) to the situations describing what they would do to respond effectively to each situation. Several strategies were used to elicit response options, including written alternatives generated by individuals and alternatives arising out of small group discussions. The written short answers were content analyzed by research staff and more response alternatives were added. The initial set consisted of 236 situations.

Field test incumbents responded to the experimental items by assessing the effectiveness of each listed response option on a scale of 1 to 7, and by indicating which option they believed was most and which least effective. During the analysis of the field test data, the content of open-ended responses from higher rated versus lower rated soldiers was compared to help guide the generation of more response alternatives. In addition, the perceived effectiveness levels (i.e., effectiveness ratings) of response alternatives from higher rated soldiers were compared with those from lower rated soldiers. Response alternatives were revised and some situations dropped between the first and second field tests. Similar comparisons and revisions were carried out between the second and third field tests.

Two additional workshops were then conducted at Fort Devens and Fort Sam Houston, with seven to nine NCOs in each. At these workshops effectiveness scale values were gathered from "expert" NCOs for each response alternative, the SJT was revised and refined, and a scoring key was developed.

A final set of 35 test items was selected on the basis of four criteria: (a) good agreement among SMEs on "correct" responses, less agreement among incumbents; (b) item content representation; (c) good distractors; and (d) proponent feedback from the Sergeants Major Academy (USASMA). There are three to five response options per item. Examinees are asked to indicate the most and least effective response alternative to each situation. The Reading Grade

Level of the test, as assessed using the FOG index (Gunning, 1952), is seventh grade.

In addition to supplying SMEs to generate scaling data, the USASMA provided a proponent review of the final test. As was true also for the supervisory simulation exercises, reviewers from USASMA considered the SJT to be a fair and appropriate method for assessing supervisory performance. The SJT also shares with the role-plays (see below) the limitation that the final version was not fully field tested before being administered to the CVII sample. Consequently, the CVII data collection is most appropriately considered a field test.

Supervisory Simulation Exercises

Three role-play simulations of supervisory behavior were developed:

- o Counseling of a subordinate with personal problems.
- o Counseling of a subordinate with performance problems.
- o Remedial training with a subordinate.

These particular simulations were developed because they cover three of the most critical tasks in the supervisory component of the NCO job, as identified in the job analyses.

The general format for the simulations is for the examinee to play the role of a supervisor. The examinee is prepared for the role with a one-page description of the situation that he or she will be asked to handle. The subordinate is played by a confederate who is trained to act out a detailed role, and who also is responsible for scoring the performance of the supervisor (i.e., examinee).

The initial developmental steps involved drafting four documents: (a) a description of the supervisor's role, (b) a short description of the subordinate's role, (c) a set of detailed instructions for playing the part of the subordinate, and (d) a performance rating instrument. Project staff drafted a checklist of behaviors applicable to performance in a counseling situation, to be used as a rating device. This checklist was generated using NCO instructional materials provided by the Army.

Participants in subsequent pilot tests tried out the role plays and provided input for refining them. This was an iterative process with participants in the later pilot tests trying out simulation materials that had already gone through several revisions. These tryouts involved considerable shadow-scoring (i.e., scoring by a second scorer) as a means of evaluating the reliability of the rating checklist. During this time the performance checklist evolved into a rating scale format. Anchors for three possible ratings were developed for each performance behavior. The simulation exercises and the rating scales used to score them are described in more detail later in this chapter.

The plan for administering the simulation exercises to the second-tour personnel in the CVII sample involved the use of civilians, hired and trained specifically for this data collection, as the role-play confederates. It was decided that the most suitable role-player candidates would be young men with

prior military experience. Once hired, role players were given at least 3 days of training in a centralized location.

Before being administered to the validation sample, the simulation exercise materials were submitted to the USASMA for a proponent review. These reviewers found the exercises to be an appropriate and fair assessment of supervisory skills, and did not request any revisions. At this point, the supervisory simulations were deemed ready for administration to the CVII sample.

Supplemental Information

Several instruments designed to obtain supplemental information were included in the set of second-tour measures:

Army Job Satisfaction Questionnaire. The Army Job Satisfaction Questionnaire was administered to both first-tour and second-tour soldiers.

Job History Questionnaire. A job history questionnaire was included in the final set of second-tour criterion measures. This instrument is the same as that used for first-tour soldiers except that it lists the tasks selected for second-tour soldier testing.

Background Information Form. As with the first-tour soldiers, it was necessary to gather a few items of descriptive information on each examinee (e.g., Social Security Number). The Background Information Form developed for second-tour soldiers also included several questions related to the extent of the examinee's supervisory experience.

Measurement Method Rating. Because two novel testing strategies were to be incorporated into the set of second-tour criterion measures, a Measurement Method Rating form was also included. This form is similar to the one used during the Concurrent Validation, but was modified to reflect the new testing methods.

The complete array of second-tour measures and supplemental information forms is listed in Table 5.1.

SECOND-TOUR DATA COLLECTION

The LVI and CVII data were obtained concurrently as one integrated data collection. The data collection began in July 1988 and was completed in February 1989. One purpose was to test first-tour soldiers who had taken the Experimental Predictor Battery as they entered the Army (the LVI sample). A second purpose was to collect second-tour performance data (the CVII sample) and, if at all possible, include the soldiers who had also participated in the Concurrent Validation (the CVI sample).

CVII data were collected at 10 CONUS installations and USAREUR sites. The data collection schedule at those installations is shown at Table 5.2.

Table 5.1

Second-Tour Criterion Measures and Supplemental Information

Criterion Measures:

Army-Wide Performance Rating Scales II^a
MOS-Specific Rating Scales II
Combat Performance Prediction Scales
Hands-on Tests II
Job Knowledge Tests II
Personnel File Form II
Situational Judgment Test
Supervisory Simulation Exercises

Supplemental Information:

Army Job Satisfaction Questionnaire
Job History Questionnaire II
Background Information Form II
Measurement Method Rating II

^a "II" indicates that this version is for second-tour soldiers.

Table 5.2

CVII Data Collection Test Dates, 1988-89

<u>Post</u>	<u>Dates</u>
Fort Lewis	11 Jul- 5 Aug
Fort Bragg	18 Jul-17 Aug
Fort Riley	19 Jul-11 Aug
Fort Hood	25 Jul-24 Aug
Fort Ord	6 Sep-30 Sep
Fort Campbell	3 Oct-28 Oct
USAREUR	10 Oct-16 Feb
Fort Polk	17 Oct-10 Nov
Fort Carson	2 Dec-16 Dec
Fort Stewart	3 Jan- 3 Feb

Data Collection Procedures

Advance Coordination

Advance site coordination for each military installation was accomplished via extensive correspondence and either one or two test site visits. The first visit provided briefings to post commanders and/or their representatives to clarify the data collection objectives, activities, and requirements. One to two weeks before the actual data collection, project staff members visited the installation to examine the test site and discuss equipment, supplies, and other special requirements for the data collection and set-up of the hands-on test stations.

Using updated listings from the Army's Worldwide Locator Service, we gave post Points-of-Contact officers (POCs) a list of the names of target examinees who were shown to be stationed on that post. The POCs used this list to identify the soldiers whom they needed to schedule for testing. To ensure that sufficient data from each MOS were collected, the samples were augmented with soldiers who were not in the original sample but were in the appropriate MOS with the requisite time in service to make them comparable to the characteristics of the target examinees. The operational definition of second-tour soldier was any individual who first entered the service during the period 1 July 1983 to 30 June 1984.

Test Site Staffing and Training

Typically, each test site required the following personnel:

Test Site Manager (TSM)	1
Hands-on Managers (HOM)	2
Hands-on Assistants	2
Paper-and-Pencil, Rating Scale, and Role-Play Administrators	5

Additionally, the Army posts provided eight NCOs per MOS to administer and score hands-on tests.

Training of Primary Staff. Most of the nonmilitary test site staff were permanent employees of the contractor consortium. However, a substantial number of additional primary staff had to be hired on a temporary basis because of the special requirements imposed by the role play in the supervisory simulation exercises. These additional test site personnel played the roles of problem subordinates in the role-play simulations and served as the role-play scorers. Much of the training for in-house staff members took place during the Concurrent Validation and the second-tour field tests. In addition, a formal training program was conducted just prior to the start of the LVI/CVII data collection trips. The training materials were covered in a 2-day training session. The individuals who were designated role players had an additional 3 days of intensive role-play actor-scorer instruction.

Hands-On Scorer Training. Training of all military scorers at the test sites was conducted in conjunction with the actual data collection. NCO scorers for each MOS received from 1 to 2 days of hands-on test administration training prior to the test administration (one day for first-tour tests plus one day for second-tour tests, if applicable). This training was provided on an MOS-specific basis by the HOM for that MOS. It followed the procedures that had been developed for the Concurrent Validation data collection (R. Campbell, 1985).

Daily Logistics

The schedule for administering the criterion measures was arranged so that no more than two Batch A MOS (first- and/or second tour) would be assessed on a given day. All test administration sessions began in the same way. The examinees assembled and roll was taken so that a search could start for any missing personnel. A project staff member would then introduce the soldiers to the project and review the activities in which they would participate throughout the day. The Privacy Act was read aloud to the soldiers at this time. Soldiers also identified those individuals for whom they would be able to provide peer ratings. If there were 20 or more soldiers in a Batch A MOS or if both first- and second-tour examinees were present, the total group was divided appropriately into subgroups.

On days when second-tour soldiers were being tested, there was normally one group of first-tour soldiers and one group of second-tour soldiers per MOS. The general test administration plan that was used when second-tour examinees were involved is shown in Figure 5.2. The second-tour schedule differs from the first-tour schedule in that one-half of the day was devoted to a combination of 3 hours of HO testing and 1 hour of supervisory simulation exercises, and the other one-half day was devoted to a somewhat different combination of written tests and ratings. Specifically, the time devoted to the Job Knowledge Test was reduced from 2 hours to 1 hour to make time for the 1-hour Situational Judgment Test.

It was expected that a significant percentage of second-tour soldiers would not be able to provide peer ratings. Soldiers at this level often work much more autonomously than their first-tour counterparts. Also, second-tour soldiers were tested in very small groups, thus decreasing the likelihood that there were many pairs of co-workers. To make effective use of the available time, the temperament/biographical inventory developed in Project A (the ABLE) was administered to all second-tour soldiers not scheduled to make peer ratings.

On supervisor ratings, the goal was to obtain two ratings for each examinee. Supervisor raters were identified with the assistance of the examinees and the NCO support staff. One of the project staff was responsible for coordinating efforts to (a) identify the supervisor, (b) schedule rating administration sessions with them, and (c) administer the supervisory rating sessions. The supervisory rating sessions ran concurrently with the other data collection and scorer training activities. Supervisors were requested to report on the same day as their subordinates.

<u>Time</u>	<u>1st-Tour MOS A</u>	<u>2nd-Tour MOS A</u>	<u>1st-Tour MOS B</u>	<u>2nd-Tour MOS B</u>
0730	In-Processing		In-Processing	
0800	H0	JK	JK	H0
0900	H0	X2	JK	H0
1000	H0	X2	X1	H0
1100	H0	S	X1	H0
1200	Lunch		Lunch	
1300	JK	H0	H0	JK
1400	JK	H0	H0	X2
1500	X1	H0	H0	X2
1600	X1	HOM	H0	SM

Legend:

- H0 = Hands-on Tests
- JK = Job Knowledge Test
- S = Situational Judgment Test
- X1 = Personnel File Form
 - Job History Questionnaire
 - Job Satisfaction Questionnaire
 - Peer Ratings (AW/MOS-specific BARS & Combat Scales)
 - Physical Requirements Survey
- X2 = Personnel File Information Form
 - Job History Questionnaire
 - Job Satisfaction Questionnaire
 - Peer Ratings (AW/MOS-specific BARS & Combat Scales) or ABLE
- M = Measurement Method Ratings

Note: This schedule assumes four groups of examinees (maximum n = 20);
two groups (one first-tour, one second-tour) for each of two MOS.

Figure 5.2. Batch A MOS first-/second-tour criterion administration schedule.

Assessment of Interscorer Agreement (Hands-on and Simulation Exercises)

Shadow-scoring efforts were incorporated into the LVI/CVII data collection. Interrater reliability estimation efforts focused on the first-tour hands-on tests for two Batch A MOS (11B and 91A). Four extra scorers were designated as shadow-scorers, and they followed a randomly selected subset of examinees from station to station. Thus, for a subset of 11B and 91A examinees, performance on all of their H0 tests was rated by two scorers.

Shadow-scoring data for the supervisory simulations were also collected at test locations in USAREUR. This was possible because there were always at least four trained role-players at each of these test sites and only three simulations were being conducted at any one time. Thus, one individual was available to observe one of the ongoing simulations and provide an independent set of scores for the examinee. Again, the issue was whether the performance ratings assigned by the role-player scorers are reliable across different scorers.

Sample Sizes

Pending completion of data editing for each criterion measure, exact sample sizes cannot be specified. However, Table 5.3 provides reasonable estimates of the number of CVII soldiers for whom a data record was established. The frequencies for each criterion measure will depend on the extent of missing data for that particular measure. This information is discussed in more detail in later sections of this report that describe the analyses for the various measures.

Table 5.3

CVII Data Collection Totals

Second-Tour Soldiers											
Post	11B	13B	19E	19K	31C	63B	71L	88M	91A	95B	Total
Lewis	19	14	8	-	12	17	17	17	14	9	127
Riley	-	14	-	-	5	11	7	14	5	16	72
Bragg	13	18	-	-	11	9	11	13	11	-	86
Hood	-	13	-	-	15	11	12	14	8	-	73
Ord	9	9	-	-	5	8	7	6	6	7	57
Campbell	21	18	-	-	12	10	9	15	15	10	110
USAREUR	28	32	-	-	31	19	38	52	28	56	284
Polk	15	13	-	10	5	13	7	13	6	15	97
Carson	18	16	25	-	-	11	-	-	-	16	86
Stewart	4	15	-	-	7	7	4	-	12	12	61
Total	127	162	33	10	103	116	112	144	105	141	1053

ANALYSIS FOR CVII BASIC CRITERION SCORES

Using the data provided by the CVII data collection, a series of analyses were carried out to establish the basic scoring procedure for each of the second-tour criterion measures. The procedure and results for each measure are described below. The end product is the set of basic scores that will enter the analysis for modeling second-tour performance (see Chapter 6).

CVII Army-Wide and Mos-Specific Rating Scales

This section reports results of the analyses of the second-tour performance. There were two major objectives: (a) to evaluate the psychometric properties of the ratings to ensure that they were of high quality, and (b) identify a set of rating scale composites (basic scores) that would appropriately reflect the performance rating content.

Content for Second-Tour Rating Scales

As reported previously, first-tour Army-wide and MOS-specific scales were modified to reflect the somewhat different job performance requirements and increased supervisory responsibilities of the second-tour soldier. For some of the rating dimensions this required few changes; for others, changes were more extensive.

For the Army-wide scales, this development effort resulted in the behavior-based first-tour Leadership dimension being replaced by three more specific leadership dimensions: Supervising, Training/Development, and Consideration for Subordinates. Further, as a result of the task-based job analysis of the leadership and supervisory responsibilities of the second-tour soldier job, seven additional scales were developed: Acting as a Role Model for Subordinates, Communication, Personal Counseling, Monitoring Subordinate Performance, Organizing Missions/ Operations, Personnel Administration, and Performance Counseling/Correcting.

For the MOS-specific scales targeted toward each of the nine Batch A MOS, the technically oriented dimensions remained substantially the same as the first-tour versions. However, some behavioral anchors were altered to reflect the increased skill expectations and additional responsibilities of the second-tour job. In five of the nine MOS, one (two in the case of 11B) MOS-specific leadership dimension was added, as well (e.g., Leading the Team for 11B). These scale development activities are reported in Campbell (1989).

In the field test of the new second-tour Army-wide and MOS-specific rating scales, supervisors and peers of approximately 250 second-tour soldiers in the nine Batch A MOS were trained to use the rating scales and then evaluated these soldiers using the scales. Results of the field test indicated (a) for several of the MOS, ratees had few peers who were qualified to rate them; (b) supervisor and peer ratings obtained showed reasonable variance across ratees; and (c) levels of interrater reliability within rater source (i.e., supervisor and peer) and across source were also acceptable and comparable to those found for first-tour soldier ratees. Minor revisions were made to the rating scale administration procedures based on field test

experience, but no changes were made to Army-wide and MOS-specific scales for the main CVII administration.

After analysis of the field test data and revision of the administration procedures, the second-tour rating scales were designated as ready for the CVII data collection. The general characteristics of the CVII sample were described in the preceding section. The specific features of the sample and data collection that are relevant for analysis of the CVII rating scale data are described below.

Sample and Data Collection Procedures

The target sample was the group of 1,053 second-tour soldiers for whom we sought to obtain supervisor and peer ratings. Data collection administrators and the Army POCs identified and contacted supervisors and peers of the target sample members. Peer ratings were generally obtained only when peers of the ratees were also members of the target sample. In a few cases, when no supervisor ratings could be obtained, peers who were not members of the target sample but who knew the work of soldiers in the sample were brought in to make peer ratings. Supervisor and peer rating sessions were usually conducted separately.

Table 5.4 shows, by MOS, the actual number of supervisor and peer raters who provided ratings for each member of the target sample. Across all MOS, more than half of the target ratees (587 of 1,053) had no peer ratings; more than 80 percent (859 of 1,053) obtained at least one supervisor rating. For those ratees who had at least one peer rating, an average of only 1.75 peers per ratee made ratings. For the ratees who received supervisory ratings, there were an average of 1.85 supervisor raters per ratee.

An extremely important aspect of each rating session was a rater orientation and training program developed to reduce various rating errors (e.g., halo) and to persuade raters to provide evaluations that were as accurate as possible. The orientation/training program was an adaptation of the program developed for raters participating in the first-tour data collection (Pulakos & Borman, 1986).

Data Analysis Plan

Analyses first examined distributions (e.g., means and standard deviations) and interrater reliabilities. Analyses for the Army-wide ratings were carried out on the total sample; MOS-specific ratings were of course analyzed separately by MOS.

Principal factor analyses with varimax rotation were conducted on the Army-wide ratings (across all MOS), for supervisor and peer ratings separately and pooled together. Because of the small Ns for individual MOS, the within-MOS factor analyses (using the same method) which we attempted on the pooled peer and supervisor ratings for the four MOS with more than 100 ratees were exploratory in nature.

Table 5.4

Number of CVII Peer and Supervisor Ratings Per Ratee by MOS

	MOS									Total Sample
	118	138	19E	31C	63B	71L	88M	91A	95B	
PEERS										
Number of Ratings										
0	70	83	17	73	89	91	42	79	43	587
1	24	49	10	13	25	18	24	20	22	205
2	21	20	12	9	2	3	37	4	24	132
3	8	6	3	6	0	0	21	1	24	69
4	4	3	1	2	0	0	19	1	23	53
5	0	1	0	0	0	0	1	0	4	6
6	0	0	0	0	0	0	0	0	1	1
CVII Sample Total N	127	162	43	103	116	112	144	105	141	1053
Mean Number of Ratings Per Ratee										
Total CVII Sample	.83	.77	1.09	.55	.25	.21	1.68	.34	1.84	.84
Rated Soldiers Only	1.86	1.57	1.81	1.90	1.07	1.14	2.37	1.38	2.65	1.75
SUPERVISORS										
Number of Ratings										
0	29	26	3	26	27	22	23	20	18	194
1	14	23	11	16	36	31	24	25	15	195
2	71	101	27	60	49	58	83	56	78	583
3	12	11	1	1	4	1	12	4	25	71
4	1	1	1	0	0	0	2	0	5	10
CVII Sample Total N	127	162	43	103	116	112	144	105	141	1053
Mean Number of Ratings Per Ratee										
Total CVII Sample	1.54	1.62	1.67	1.35	1.26	1.34	1.63	1.42	1.89	1.52
Rated Soldiers Only	2.00	1.93	1.80	1.81	1.64	1.67	1.93	1.75	2.16	1.85

Results and Discussion

For each type of ratings data (Army-wide and MOS-specific), results describing the distributional properties are presented first. Second, interrater reliability results are shown. Finally, factor analysis results are summarized.

Army-Wide Scales: Rating Distributions. The Army-wide rating distributions are shown in Tables 5.5 and 5.6. These results demonstrate that raters used all scale points, although "1" and "2" were infrequent, especially for peer raters. In addition, peers and supervisors both seemed to provide lower ratings on the supervisory dimensions compared to the non-supervisory dimensions.

These results are supported by Table 5.7 as well. The overall mean for peer raters on the supervisory-oriented dimensions is 4.72, versus 5.08 for non-supervisory dimensions. The corresponding supervisor rater overall mean ratings are 4.50 and 5.08, respectively. Thus, supervisors provided slightly lower mean ratings on the supervisory dimensions and exactly the same level mean ratings on the non-supervisory dimensions, compared to their peer rater counterparts. However, this comparison should be viewed with some caution, because the data for peer and supervisor ratings are derived from substantially different samples of ratees.

Table 5.7 also indicates that both supervisor and peer ratings of second-tour soldiers on the non-supervisory dimensions were about half a scale point higher than the corresponding ratings of first-tour soldiers in the CVI research. A possible hypothesis would be that second-tour personnel should perform at a somewhat higher level on the technical part of the job compared to their first-tour counterparts. For those rating dimensions where the first-tour anchors had been modified in developing the second-tour rating scale, the differences obtained are probably underestimates of the true differences because the anchors used to make the second-tour ratings reflected higher performance requirements than those used to make the first-tour ratings.

The differentiation among ratees is indicated by the standard deviations in Table 5.7. On average, the peer ratings provide more differentiation for second-tour soldiers than was the case for the first-tour ratings. The supervisory ratings of second-tour performance have about the same variance as the ratings of first-tour performance. In addition, the standard deviations of the supervisor and peer ratings for second-tour soldier performance are quite comparable.

Overall, the distributions of supervisor and peer ratings of second-tour soldier performance on the Army-wide scales seem appropriate. They show few signs of errors of central tendency or leniency.

Army-Wide Scales: Interrater Reliability. Interrater reliability results for the Army-wide scales are presented in Tables 5.8 and 5.9. These are shown for peer, supervisor, and pooled peer and supervisory ratings. Table 5.8 contains intraclass correlations that reflect the reliability of a single rater. Table 5.9 indicates the reliability of the mean peer, supervisor, and pooled peer/supervisor ratings, respectively. Since these estimates are a function of the average number of raters per ratee, the

Table 5.5

CVII Army-Wide Ratings: Use of Scale Points by Peers (Percent)

	7	6	5	4	3	2	1
<i>Behavior Scales^a</i>							
1. Technical Knowledge/Skill*	11	36	27	16	7	3	0
2. Effort*	12	29	29	15	10	4	1
3. Supervising	8	23	28	19	15	5	2
4. Follow Regs./Orders*	14	29	28	14	9	4	2
5. Integrity*	14	31	27	15	7	4	2
6. Training/Developing	9	24	27	21	11	6	2
7. Maintain Equipment*	14	29	30	18	6	2	1
8. Physical Fitness*	19	23	27	15	10	4	2
9. Self-Development*	12	24	29	18	10	6	1
10. Consideration for Subord.	11	29	28	18	9	4	1
11. Military Bearing*	19	28	26	15	7	4	1
12. Self-Control*	16	26	25	16	10	6	1
<i>Additional Leadership Scales</i>							
13. Role Model	7	21	28	23	11	7	3
14. Communication	10	22	33	17	12	5	1
15. Personal Counseling	7	19	28	23	15	5	3
16. Monitoring	9	23	32	19	10	4	3
17. Organize Missions/Operations	9	21	32	20	12	4	2
18. Personnel Administration	10	22	26	20	15	4	3
19. Performance Counseling	10	20	32	21	11	4	2
20. Overall Effectiveness	10	30	33	16	6	4	1
21. Senior NCO Potential	16	27	26	15	8	4	4

Mean Non-Supervisory*	14.56	26.11	27.56	15.78	8.44	4.11	1.22
Mean Supervisory	9.00	20.60	29.40	20.10	12.10	4.80	2.90

Note. Sample sizes range from 974 to 989 for the behavior scales and from 918 to 962 for the additional leadership scales.

^aAn asterisk designates non-supervisory scales.

Table 5.6

CVII Army-Wide Ratings: Use of Scale Points by Supervisors (Percent)

	7	6	5	4	3	2	1
<i>Behavior Scales^a</i>							
1. Technical Knowledge/Skill*	11	31	29	18	8	3	0
2. Effort*	14	25	27	17	10	6	1
3. Supervising	6	16	24	22	19	11	2
4. Follow Regs./Orders*	15	27	29	13	8	6	2
5. Integrity*	20	28	24	13	7	6	2
6. Training/Development	6	18	27	23	14	11	1
7. Maintain Equipment*	15	23	30	17	10	4	1
8. Physical Fitness*	23	22	26	13	9	5	2
9. Self-Development*	10	21	29	19	13	6	2
10. Consideration for Subord.	12	25	30	17	11	4	1
11. Military Bearing*	21	23	27	15	9	4	1
12. Self-Control*	23	27	22	14	8	5	1
<i>Additional Leadership Scales</i>							
13. Role Model	7	18	28	20	16	9	2
14. Communication	7	21	30	22	14	5	1
15. Personal Counseling	5	15	28	23	18	9	2
16. Monitoring	6	18	30	24	13	8	1
17. Organize Missions/Operations	7	18	28	22	16	7	2
18. Personnel Administration	7	16	28	25	14	8	2
19. Performance Counseling	6	15	29	22	16	9	3
20. Overall Effectiveness	7	24	35	18	11	5	0
21. Senior NCO Potential	12	24	24	16	12	9	3

Mean Non-Supervisory*	16.89	25.22	27.00	15.44	9.11	5.00	1.33
Mean Supervisory	6.90	18.00	28.20	22.00	15.10	8.10	1.70

Note. Sample sizes range from 1602 to 1732 for the behavior scales and from 1502 to 1654 for the additional leadership scales.

^aAn asterisk designates non-supervisory scales.

Table 5.7

CVII Army-Wide Ratings: Means and Standard Deviations

	Peers		Supervisors	
	Mean	SD	Mean	SD
<i>Behavior Scales^a</i>				
1. Technical Knowledge/Skill*	5.16	1.05	5.13	1.07
2. Effort*	5.04	1.17	4.97	1.25
3. Supervising	4.64	1.28	4.27	1.29
4. Follow Regs./Orders*	5.08	1.20	5.04	1.23
5. Integrity*	5.10	1.22	5.18	1.29
6. Training/Development	4.71	1.24	4.43	1.27
7. Maintain Equipment*	5.18	1.08	5.03	1.20
8. Physical Fitness*	5.08	1.31	5.17	1.38
9. Self-Development*	4.83	1.21	4.74	1.24
10. Consideration for Subord.	4.99	1.17	4.95	1.18
11. Military Bearing*	5.18	1.21	5.19	1.26
12. Self-Control*	5.03	1.24	5.25	1.30
<i>Additional Leadership Scales</i>				
13. Role Model	4.60	1.24	4.46	1.28
14. Communication	4.78	1.17	4.66	1.14
15. Personal Counseling	4.49	1.27	4.32	1.26
16. Monitoring	4.77	1.22	4.52	1.18
17. Organize Missions/Operations	4.76	1.19	4.53	1.21
18. Personnel Administration	4.70	1.28	4.47	1.23
19. Performance Counseling	4.80	1.15	4.36	1.24
20. Overall Effectiveness	5.05	1.12	4.83	1.10
21. Senior NCO Potential	5.03	1.30	4.72	1.41

Mean Across Non-Supervisory Dimensions*	5.08	1.19	5.08	1.25
Mean Across Supervisory Dimensions	4.72	1.22	4.50	1.23
Mean Across CVI Batch A Rating Dimensions (Peer N = 4,902 ratees, 15,985 raters; Supervisor N = 4,943 ratees, 9,392 raters.)	4.60	1.02	4.54	1.30

Note. Peers, N = 484-500; Supervisors, N = 857-927.

^aAn asterisk indicates non-supervisory scales.

Table 5.8

CVII Army-Wide Ratings: One-Rater Interrater Reliability

	Peer Ratings	Supervisor Ratings	Pooled Peer/Sup. Ratings
<i>Behavior Scales</i>			
1. Technical Knowledge/Skill	.19	.41	.30
2. Effort	.16	.40	.30
3. Supervising	.31	.36	.30
4. Follow Regs./Orders	.16	.36	.26
5. Integrity	.17	.35	.24
6. Training/Development	.27	.37	.28
7. Maintain Equipment	.14	.29	.20
8. Physical Fitness	.35	.53	.45
9. Self-Development	.20	.36	.27
10. Consideration for Subord.	.20	.36	.23
11. Military Bearing	.23	.46	.35
12. Self-Control	.16	.33	.25
<i>Additional Leadership Scales</i>			
13. Role Model	.31	.44	.34
14. Communication	.21	.33	.23
15. Personal Counseling	.23	.39	.27
16. Monitoring	.23	.33	.26
17. Organize Missions/Operations	.19	.33	.25
18. Personnel Administration	.19	.32	.24
19. Performance Counseling	.15	.33	.24
20. Overall Effectiveness	.25	.41	.30
21. Senior NCO Potential	.20	.45	.31

Median for Behavior Scales	.20	.36	.28
Median for Additional Leadership Scales	.21	.33	.25
Average Ratings Per Ratee	1.99	1.88	2.77
Median for CVI Behavior Scales	.22	.37	---

Note. The total number of ratings used to compute reliabilities ranged from 918 to 989 for peers, from 1495 to 1735 for supervisors, and from 2415 to 2720 for pooled peers/supervisors.

Table 5.9

CVII Army-Wide Ratings: k -Rater Interrater Reliability^a

	Peer Ratings	Supervisor Ratings	Pooled Peer/Sup. Ratings
<i>Behavior Scales</i>			
1. Technical Knowledge/Skill	.32	.57	.54
2. Effort	.27	.56	.55
3. Supervising	.47	.51	.54
4. Follow Regs./Orders	.28	.51	.49
5. Integrity	.28	.51	.47
6. Training/Development	.42	.52	.52
7. Maintain Equipment	.24	.43	.41
8. Physical Fitness	.51	.68	.69
9. Self-Development	.33	.51	.51
10. Consideration for Subord.	.32	.51	.46
11. Military Bearing	.38	.62	.61
12. Self-Control	.28	.48	.48
<i>Additional Leadership Scales</i>			
13. Role Model	.47	.59	.59
14. Communication	.34	.47	.46
15. Personal Counseling	.36	.54	.49
16. Monitoring	.36	.47	.49
17. Organize Missions/Operations	.31	.47	.49
18. Personnel Administration	.31	.45	.47
19. Performance Counseling	.26	.47	.46
20. Overall Effectiveness	.39	.56	.55
21. Senior NCO Potential	.33	.60	.55

Median for Behavior Scales	.32	.51	.52
Median for Additional Leadership Scales	.34	.47	.49
Average Ratings Per Ratee	1.99	1.88	2.77
Median for CVI Behavior Scales	.48	.52	---

Note. The total number of ratings used to compute reliabilities ranged from 918 to 989 for peers, from 1495 to 1735 for supervisors, and from 2415 to 2720 for pooled peers/supervisors.

^a k is the mean number of ratings per ratee.

one-rater indexes in Table 5.8 are best for comparing the relative levels of interrater agreement for different rater groups.

First, Table 5.8 shows that the degree of interrater agreement for peers and supervisors is almost exactly the same as was found for the CVI sample. Second, supervisors again provide more reliable ratings than do peers. Third, the pooled peer and supervisor ratings are actually less reliable than the supervisor ratings by themselves. Finally, the supplemental leadership scales with a less behavioral format have about the same reliability as the behavior-based rating scales.

Table 5.9 indicates that the mean ratings are reasonably reliable, especially the supervisor ratings. Compared to CVI, the mean peer ratings in the present sample have lower reliability levels, which is due to the smaller number of peers per ratee in the second-tour sample. When peer and supervisor ratings are pooled, the additional numbers of raters per ratee help to bring the level of reliability for the combined peer/supervisor ratings to about the same level as the supervisor ratings by themselves.

In sum, the interrater reliabilities in the present sample are very nearly the same as were found in CVI. The mean supervisor ratings and the pooled peer and supervisor ratings have very acceptable levels of reliability. However, in terms of their reliability and distributional properties the supervisor ratings are superior to the peer ratings for this sample.

Army-Wide Scales: Factor Analysis Results. Several factor analyses were conducted on the second-tour Army-wide soldier ratings from the CVII sample. Army-wide ratings on the nine non-supervisory second-tour dimensions were intercorrelated and factor analyzed so that the CVI and CVII factor structures could be compared for these non-supervisory dimensions. Then, the ratings on the 10 supervisory dimensions for the CVII sample were intercorrelated and factor analyzed to assess the possibility of multiple underlying leadership/supervision factors. Finally, the same procedure was followed for all 19 of the Army-wide dimensions.

Table 5.10 demonstrates the remarkable similarity of the rotated factor structures for the nine non-supervisory dimensions that are common to the first- and second-tour ratings scales. The three factors obtained in the CVI sample were closely replicated with the CVII data.

Factor analysis of the 10 supervisory dimensions resulted in a single leadership/supervision factor. Consequently, these results are not presented.

Tables 5.11-5.13 show the four-factor rotated solutions for, respectively, the peer, supervisor, and pooled peer/supervisor ratings. The peer solution is not quite as clear as the supervisor or peer/supervisor solutions; the latter two solutions have three factors very similar to the CVI factors and a separate leadership/supervision factor. The lack of clarity in the peer solution is partly a function of the lower reliabilities and somewhat poorer distributional properties.

Parallel to the preceding discussion of the Army-wide scales, results relative to MOS-specific rating distributions, interrater reliabilities, and factor structures are described below.

Table 5.10

Comparison of CVI and CVII Factor Analyses^a: Pooled Peer/Supervisor Ratings^b,
Non-Supervisory Dimensions Only

Dimension	Factor Loadings (CVI/CVII)			h ² ^c
	1	2	3	
Technical Knowledge/Skill	<u>.71/.71</u>	.28/.28	.30/.27	.67/.66
Leadership	<u>.69/--</u>	.30/--	.37/--	.70/--
Effort	<u>.69/.73</u>	.43/.36	.26/.29	.73/.75
Self-Development	<u>.57/.56</u>	.38/.31	.38/.49	.61/.65
Maintain Equipment	<u>.54/.52</u>	.34/.36	.35/.33	.53/.51
Follow Regulations	.41/.42	<u>.69/.66</u>	.30/.33	.73/.72
Self-Control	.22/.18	<u>.63/.57</u>	.20/.19	.49/.39
Integrity	.50/.45	<u>.59/.67</u>	.28/.26	.68/.72
Military Bearing	.32/.31	.32/.38	<u>.57/.62</u>	.53/.62
Physical Fitness	.21/.23	.15/.18	<u>.49/.62</u>	.31/.47

Eigenvalue	2.69/2.18	1.96/1.82	1.33/1.49	5.98/5.49

Note. Sample size is 9845 for CVI and 950 for CVII.

^aPrincipal factor analysis, varimax rotation.

^bComputed by averaging the mean peer rating and the mean supervisor rating.

^ch² = communality (sum of squared factor loadings) for variables.

Table 5.11

CVII Army-Wide Factor Analysis^a: Peer Ratings, All Dimensions

Dimension	Factor Loadings				h^2 ^b
	1	2	3	4	
1. Technical Knowledge/Skill	.38	<u>.57</u>	.32	.20	.61
2. Effort	.42	<u>.58</u>	.21	.27	.63
3. Supervising	.55	.53	.30	.16	.70
4. Follow Regs./Orders	.29	.40	.45	<u>.50</u>	.70
5. Integrity	.30	.44	.29	<u>.53</u>	.65
6. Training/Development	<u>.55</u>	.43	.30	.21	.62
7. Maintain Equipment	.28	.39	.20	.36	.40
8. Physical Fitness	.21	.28	<u>.56</u>	.26	.50
9. Self-Development	.45	.46	.38	.24	.61
10. Consideration for Subord.	.43	.38	.31	.38	.57
11. Military Bearing	.31	.24	<u>.63</u>	.30	.55
12. Self-Control	.27	.10	.31	<u>.48</u>	.41
13. Role Model	<u>.56</u>	.35	.43	.36	.75
14. Communication	<u>.62</u>	.29	.19	.35	.63
15. Personal Counseling	<u>.69</u>	.34	.18	.24	.68
16. Monitoring	<u>.68</u>	.31	.31	.25	.72
17. Organize Missions/ Operations	<u>.71</u>	.22	.26	.24	.68
18. Personnel Administration	<u>.66</u>	.26	.16	.27	.60
19. Performance Counseling	<u>.66</u>	.26	.32	.18	.64

Eigenvalue	4.79	2.74	2.25	1.97	11.65

Note. Sample size is 473.

^aPrincipal factor analysis, varimax rotation.

^b h^2 = communality (sum of squared factor loadings) for variables.

Table 5.12

CVII Army-Wide Factor Analysis^a: Supervisor Ratings, All Dimensions

Dimension	Factor Loadings				h^2 ^b
	1	2	3	4	
1. Technical Knowledge/Skill	.41	<u>.65</u>	.24	.22	.70
2. Effort	.39	<u>.68</u>	.31	.27	.78
3. Supervising	.57	.53	.21	.28	.73
4. Follow Regs./Orders	.29	.36	<u>.63</u>	.30	.70
5. Integrity	.32	.37	<u>.66</u>	.22	.72
6. Training/Development	.52	.52	.24	.27	.67
7. Maintain Equipment	.32	<u>.50</u>	.33	.25	.52
8. Physical Fitness	.20	.19	.18	<u>.60</u>	.47
9. Self-Development	.41	.48	.27	.44	.67
10. Consideration for Subord.	.47	.40	.44	.26	.64
11. Military Bearing	.30	.22	.34	<u>.63</u>	.65
12. Self-Control	.17	.09	<u>.56</u>	.18	.38
13. Role Model	.53	.31	.40	.51	.80
14. Communication	<u>.62</u>	.35	.34	.23	.68
15. Personal Counseling	<u>.72</u>	.19	.31	.26	.72
16. Monitoring	<u>.63</u>	.41	.31	.22	.71
17. Organize Missions/ Operations	<u>.70</u>	.36	.26	.20	.73
18. Personnel Administration	<u>.63</u>	.29	.20	.24	.58
19. Performance Counseling	<u>.72</u>	.32	.20	.29	.74

Eigenvalue	4.74	3.17	2.54	2.17	12.54

Note. Sample size is 823.

^aPrincipal factor analysis, varimax rotation.

^b h^2 = communality (sum of squared factor loadings) for variables.

Table 5.13

CVII Army-Wide Factor Analysis^a: Pooled Peer/Supervisor Ratings^b, All Dimensions

Dimension	Factor Loadings				h^2 ^c
	1	2	3	4	
1. Technical Knowledge/Skill	.45	.26	<u>.60</u>	.22	.68
2. Effort	.44	.35	<u>.61</u>	.26	.76
3. Supervising	<u>.62</u>	.20	.48	.30	.74
4. Follow Regs./Orders	.32	<u>.63</u>	.32	.32	.70
5. Integrity	.33	<u>.66</u>	.34	.23	.71
6. Training/Development	<u>.58</u>	.23	.46	.28	.68
7. Maintain Equipment	.35	.36	<u>.42</u>	.28	.51
8. Physical Fitness	.21	.18	.16	<u>.61</u>	.47
9. Self-Development	.46	.27	.43	.44	.66
10. Consideration for Subord.	<u>.51</u>	.44	.35	.27	.65
11. Military Bearing	.29	.36	.21	<u>.64</u>	.67
12. Self-Control	.18	<u>.59</u>	.08	.18	.42
13. Role Model	.54	.43	.26	.51	.80
14. Communication	<u>.62</u>	.37	.29	.23	.66
15. Personal Counseling	<u>.73</u>	.29	.16	.25	.71
16. Monitoring	<u>.67</u>	.32	.33	.24	.72
17. Organize Missions/ Operations	<u>.73</u>	.26	.29	.21	.73
18. Personnel Administration	<u>.64</u>	.23	.26	.20	.57
19. Performance Counseling	<u>.73</u>	.22	.26	.30	.74

Eigenvalue	5.22	2.70	2.46	2.21	12.58

Note. Sample size is 902.

^aPrincipal factor analysis, varimax rotation.

^bComputed by averaging the mean peer rating and the mean supervisor rating.

^c h^2 = communality (sum of squared factor loadings) for variables.

MOS-Specific Scales: Rating Distributions. The means and standard deviations of the MOS-specific ratings for each MOS are presented in Table 5.14. Peer and supervisor rating results are shown separately, as are the supervisory- and non-supervisory-oriented dimensions. In general, the means and standard deviations of the ratings are larger for this CVII sample than they were for the first-tour soldier. Also, the means for the MOS-specific non-supervisory dimensions, peer and supervisor ratings, are higher than the Army-wide non-supervisory dimensions in this CVII sample. The unweighted mean across MOS of the MOS-specific ratings is 5.26 for the peers and 5.24 for the supervisors, whereas the Army-wide CVII means were 5.08 for both rating sources.

MOS-Specific Scales: Interrater Reliability. Interrater reliability results are shown in Tables 5.15 and 5.16. Table 5.15 presents one-rater reliabilities. As with the Army-wide ratings, peer ratings are not very reliable at the one-rater level, supervisor ratings are considerably more reliable, and the pooled peer and supervisor ratings yield reliabilities at an intermediate level.

Reliabilities of the mean ratings are of course higher. However, for the mean peer ratings they are very low or near zero for three of the MOS, and equal to zero for one MOS even when the composite across all dimensions is computed. At least in part, this is because there are so few ratings per ratee in many of the MOS, and the estimates of reliability are very unstable. While the mean supervisor ratings on individual dimensions are considerably more reliable, on average they are not as reliable as the Army-wide ratings provided by supervisors. Finally, the pooled peer/supervisor ratings are actually, for the most part, less reliable than the supervisor ratings, even with the larger number of raters per ratee for the pooled ratings. Strictly speaking, peer and supervisory ratings are not parallel measures. However, they both may be valid measures of somewhat different things.

MOS-Specific Scales: Factor Analyses Results. Factor analyses of MOS-specific rating data did not yield any interpretable multiple factor solutions for any of the four individual MOS analyzed (the four MOS having more than 100 ratees). In each case, the most reasonable solution was simply a single performance factor. Therefore, none of these solutions are presented.

Summary of Rating Scale Analyses

For both the Army-wide and MOS-specific rating scales, the mean, variance, and reliability of the supervisor and pooled peer/supervisor ratings appear quite acceptable and comparable to what was found in the CVI research. Factor analyses of the Army-wide ratings suggested that the three-factor CVI solution could be replicated in the present data. A fourth Leading/Supervising factor was discernible, although it was not as distinct as the other three factors. Accordingly, the four composites shown in Table 5.17 are proposed as the basic scores for the Army-wide rating data. As in CVI, unit weighting of each dimension is recommended when computing scores for each rating composite. Definitions for each of the four composites are presented in Table 5.18.

Table 5.14

CVII MOS-Specific Ratings: Means^a Across Rating Dimensions for Each MOS

		MOS								
		11B	13B	19E	31C	63B	71L	88M	91A	95B
PEER RATINGS										
Non-Sup. Dimensions										
Mean	4.83 (11)	5.24(10)	5.60 (8)	4.95(7)	5.00 (12)	5.80(7)	5.49(11)	5.28(9)	5.17(13)	
SD	1.15	1.29	1.07	1.08	1.30	1.38	1.09	.98	.96	
Supervisory Dimensions										
Mean	4.48 (2)	5.59 (1)	5.45 (1)	4.97(1)	--	--	--	--	4.84 (1)	
SD	1.18	1.38	1.29	.91	--	--	--	--	.99	
Number of Ratees	58-62	70-72	24-25	27-28	20-23	17	92-97	19-23	83-98	
CVI Mean	4.58 (12)	4.70(10)	4.74 (8)	4.91(6)	4.64 (11)	4.87(8)	4.75(10)	4.60(9)	4.80(11)	
CVI SD	.90	.93	.87	.95	1.01	1.03	.89	.94	.85	
SUPERVISOR RATINGS										
Non-Sup. Dimensions										
Mean	5.18 (11)	5.28(10)	5.57 (8)	5.08(7)	4.97 (12)	5.17(7)	5.40(11)	5.37(9)	5.17(13)	
SD	1.10	1.08	.94	1.19	1.15	1.19	1.14	1.07	1.06	
Supervisory Dimensions										
Mean	4.61 (2)	4.96 (1)	5.52 (1)	4.93(1)	--	--	--	--	4.87 (1)	
SD	1.23	1.38	1.05	1.35	--	--	--	--	1.15	
Number of Ratees	98-103	120-135	39-41	73-78	86-88	70-84	115-121	73-85	105-123	
CVI Mean	4.56 (12)	4.69(10)	4.72 (8)	4.71(6)	4.18(11)	4.92(8)	4.75(10)	4.49(9)	4.82(11)	
CVI SD	1.17	1.18	1.06	1.13	1.27	1.18	1.12	1.09	1.06	

^aValues in parentheses represent number of dimensions used in computing mean ratings.

Table 5.15

CVII MOS-Specific Ratings: One-Rater Interrater Reliability by MOS

		MOS									
		11B	13B	19E	31C	63B	71L	88M	91A	95B	
Number of Rating Dimensions		13	11	9	8	12	7	11	9	14	
PEER RATINGS											
Median Dimension Reliability		.00	.32	.00	.04	1.00	1.00	.13	.19	.16	
Range Across Dimensions		.00-.26	.00-.47	.00-.24	.00-.16	.66-1.00	.80-1.00	.07-.42	.00-.86	.00-.25	
Reliability of Composite ^a		.13	.41	.00	.06	.97	1.00	.16	.38	.23	
Average Ratings Per Ratee		1.75	1.61	1.76	1.93	1.04	1.14	2.08	1.33	2.46	
SUPERVISOR RATINGS											
Median Dimension Reliability		.31	.24	.31	.50	.32	.33	.31	.28	.28	
Range Across Dimensions		.18-.42	.09-.45	.00-.45	.14-.65	.13-.55	.09-.51	.22-.42	.12-.42	.09-.41	
Reliability of Composite ^a		.45	.37	.40	.74	.43	.53	.57	.34	.40	
Average Ratings Per Ratee		1.93	1.90	1.78	1.73	1.58	1.59	1.90	1.68	1.96	
POOLED PEER/SUPERVISOR RATINGS											
Median Dimension Reliability		.15	.23	.12	.29	.27	.15	.14	.21	.19	
Range Across Dimensions		.07-.31	.09-.41	.00-.32	.18-.36	.01-.42	.07-.29	.09-.23	.04-.38	.07-.30	
Reliability of Composite ^a		.26	.34	.22	.45	.34	.23	.20	.26	.26	
Average Ratings Per Ratee		2.79	2.44	2.68	2.18	1.68	1.74	3.19	1.81	3.45	

Note. The total number of ratings used to compute reliabilities for each MOS ranged from 22 to 257 for peers, from 72 to 260 for supervisors, and from 112 to 517 for pooled peers/supervisors.

^aThe composite is the average rating across all dimensions.

Table 5.16

CVII MOS-Specific Ratings: k -Rater Interrater Reliability^a by MOS

		MDS									
		11B	13B	19E	31C	63B	71L	88M	91A	95B	
Number of Rating Dimensions		13	11	9	8	12	7	11	9	14	
PEER RATINGS											
Median Dimension Reliability	.00	.43	.00	.07	1.00	1.00	1.00	.22	.23	.30	
Range Across Dimensions	.00-.37	.00-.58	.00-.36	.00-.28	.67-1.00	.82-1.00		.13-.60	.00-.87	.14-.44	
Reliability of Composite ^b	.21	.53	.00	.11	.97	1.00	1.00	.28	.45	.42	
Average Ratings Per Ratee	1.75	1.61	1.76	1.93	1.04	1.14	1.14	2.08	1.33	2.46	
SUPERVISOR RATINGS											
Median Dimension Reliability	.45	.35	.44	.62	.43	.44	.44	.45	.39	.41	
Range Across Dimensions	.30-.57	.15-.59	.00-.59	.21-.76	.19-.65	.13-.62		.35-.57	.18-.54	.15-.57	
Reliability of Composite ^b	.61	.52	.54	.83	.54	.65	.65	.57	.47	.57	
Average Ratings Per Ratee	1.93	1.90	1.78	1.73	1.58	1.59	1.59	1.90	1.68	1.96	
POOLED PEER/SUPERVISOR RATINGS											
Median Dimension Reliability	.33	.43	.27	.47	.39	.23	.23	.35	.32	.48	
Range Across Dimensions	.18-.56	.19-.64	.00-.55	.32-.55	.01-.55	.13-.42		.24-.48	.07-.52	.21-.59	
Reliability of Composite ^b	.50	.57	.43	.64	.47	.35	.35	.45	.41	.58	
Average Ratings Per Ratee	2.79	2.44	2.68	2.18	1.68	1.74	1.74	3.19	1.81	3.45	

Note. The total number of ratings used to compute reliabilities for each MOS ranged from 22 to 257 for peers, from 72 to 260 for supervisors, and from 112 to 517 for pooled peers/supervisors.

^a k is the average number of ratings per ratee.

^b The composite is the average rating across all dimensions.

Table 5.17**Composition of Proposed CVII Army-Wide Rating Composites**

Composite Name	Percent Common Variance Accounted for by Relevant Factor	Dimensions Included ^a
1. Leading/Supervising	41.5	Supervising Training/Development Consideration for Subord. Communication Personal Counseling Monitoring Organize Missions/Operations Personnel Administration Performance Counseling
2. Personal Discipline	21.5	Follow Regs./Orders Integrity Self-Control
3. Technical Skill/Effort	19.6	Technical Knowledge/Skill Effort Maintain Assigned Equipment
4. Physical Fitness/ Military Bearing	17.5	Military Bearing Physical Fitness

^aTwo dimensions were not included in any composites: Act as a Role Model and Self-Development.

Table 5.18

Definitions of Proposed CVII Army-Wide Rating Composites

Leading/Supervising:

Effectively organizing, monitoring, and, when necessary, correcting subordinates; providing proper training experiences; communicating effectively to keep subordinates and superiors informed; and providing support and help to subordinates when needed.

Personal Discipline:

Adhering to Army rules and regulations; exercising self-control; demonstrating integrity in day-to-day behavior; and not causing disciplinary problems.

Technical Skill/Effort:

Displaying technical knowledge and skill in accomplishing job tasks and completing assignments; showing conscientiousness and initiative on the job; and exerting considerable effort to get jobs and tasks done effectively.

Physical Fitness/Military Bearing:

Maintaining an appropriate military appearance and bearing and staying in good physical condition.

The interrater reliabilities of these four unit-weighted composites are shown in Table 5.19. Supervisors provide the most reliable ratings; however, the pooled peer and supervisor ratings are nearly as reliable and may be more valid in the sense that they draw on both peer and supervisor perspectives. It has been argued (Borman, 1974; Campbell, Dunnette, Lawler, & Weick, 1970) that typically in organizations peers and supervisors each have relevant and important performance information on ratees but that these two rating sources may have different information because of different roles and opportunity to observe performance-related behavior. Thus, we would expect that peer and supervisor ratings, taken together, should provide a more complete assessment than either of the rating sources individually.

Correlations among the four Army-wide unit-weighted composites are presented in Table 5.20. Although some of these correlations are quite high, our experience in CVI suggests that there should be sufficient differentiation between these CVII composites to provide multidimensional performance information.

Table 5.19

One-Rater and \bar{k} -Rater Interrater Reliability^a for Army-Wide Composites

	Leading/ Supervising	Technical Skill/ Effort	Personal Discipline	Fitness/ Bearing
<i>Peer Ratings</i>				
One-rater	.33	.20	.22	.35
\bar{k} -rater	.48	.33	.36	.52
Average Ratings Per Ratee	1.90	1.95	1.96	1.96
<i>Supervisor Ratings</i>				
One-rater	.50	.48	.45	.56
\bar{k} -rater	.64	.63	.60	.70
Average Ratings Per Ratee	1.75	1.86	1.86	1.86
<i>Pooled Peer/Supervisor Ratings^b</i>				
One-rater	.37	.34	.32	.45
\bar{k} -rater	.61	.59	.57	.70
Average Ratings Per Ratee	2.62	2.79	2.81	2.82

^a \bar{k} is the average number of ratings per ratee.

^bComputed by averaging the mean peer rating and the mean supervisor rating.

Table 5.20

Intercorrelations Among Proposed CVII Rating Composites^a for Pooled Peer/Supervisor Ratings^b

	Leading/ Supervising	Technical Skill/ Effort	Personal Discipline	Fitness/ Bearing
Technical Skill/Effort	.81	---		
Personal Discipline	.68	.69	---	
Fitness/Bearing	.60	.58	.50	---

Note. Sample sizes range from 915 to 956.

^aIntercorrelations among CVI composites based on pooled peers/supervisors ranged from .51 to .64.

^bComputed by averaging the mean peer rating and the mean supervisor rating.

For the MOS-specific ratings, a unit-weighted composite of all dimension ratings for each MOS is recommended. Factor analysis results did not indicate multiple factors in any of the MOS-specific ratings analyzed.

In sum, the Army-wide ratings provide a reliable and interpretable multidimensional depiction of four different performance areas. These composites should contribute significantly to the development of the second-tour performance model (described in Chapter 6). The MOS-specific ratings provide a single, reliable job performance composite that also represents a significant component of total performance.

Combat Performance Prediction Scales

The Combat Performance Prediction Scales were summated scales intended to evaluate performance under the types of emergency, difficult, or dangerous conditions that might be found in combat. The items on the scales depicted critical incidents applicable to such situations. Raters were instructed to indicate the likelihood that the ratee would behave in the manner described by the item.

The Combat Scales were designed to be an Army-wide, rather than an MOS-specific, criterion measure. They were administered to both first- and second-tour soldiers from all tested MOS in the 1988 Project A data collection. Because females are restricted from combat MOS, ratings were not collected for them during this data collection.

An earlier version of the Combat Performance Prediction Scales was used with the Concurrent Validation sample in 1985 (Campbell, 1986). The earlier version consisted of 40 items to be rated, as well as a rating indicating the confidence the rater had in the validity of his or her judgments. Ratings were made on a 15-point scale. Two scores were derived from this instrument: "Performing Under Adverse Conditions" and "Avoiding Mistakes."

In preparation for the 1988 data collection, a group of subject matter experts reviewed the 40-item scale to verify that all items were appropriate for second-tour, as well as first-tour, soldiers. Although all items were deemed suitable for both first- and second-tour soldiers, a number of items were dropped to reduce the length of the instrument. Items that were deleted included those that appeared to measure overall technical proficiency more than performance under poor conditions and those that exhibited the lowest interrater reliabilities in the Concurrent Validation. Some of the remaining items were revised and the 15-point rating scale was reduced to a 7-point scale.

Although Combat Scale ratings were gathered from both first- and second-tour male soldiers, only the second-tour data are reported here. Combat Scale ratings by supervisors were collected for 815 second-tour soldiers and peer ratings were collected for 447. A total of 848 soldiers had supervisor and/or peer ratings.

Different forms were used to collect Combat Scale data for first- and second-tour soldiers; one form was machine scannable and the other was not. Because five items were different on the two forms, only 14 of the 19 items on the Combat Scales (in addition to the confidence rating) were included in the data analyses reported here.

Principal components analyses were used to determine how many Combat Scale subscores should be computed. Table 5.21 shows the rotated factor pattern matrix of the two-factor solution using combined supervisor and peer ratings. Although separate analyses were run for supervisor and peer ratings, the results were essentially the same and they are not reported here. The second factor is composed of the three negatively worded items on the Combat Scales. Given that this factor is probably not substantively distinct from the first, it was decided that only one score for the Combat Scales should be computed. This score would be the sum of the 14 item ratings, with the three negatively worded items reverse-scored.

Table 5.22 lists the interrater reliability estimates for the Combat Scale score when it is computed based on data from all raters and separately for supervisors and peers. The interrater reliability of the Combat Scales is similar to that of the Army-Wide BARS. Coefficient alphas for the Combat Scales scores were .93 for the combined ratings, .93 for the supervisor ratings, and .91 for the peer ratings.

Table 5.21

Principal Components Analysis of Combat Performance Prediction Scale Ratings^a

<u>Item</u>	<u>Factor 1</u>	<u>Factor 2</u>	<u>h² b</u>
01	.78	-.16	.63
02	.74	-.23	.60
03	-.15	.82	.69
04	-.17	.83	.72
05	.78	-.25	.67
06	.76	-.28	.66
07	.75	-.29	.65
08	.71	-.19	.54
09	.82	-.27	.75
10	.84	-.22	.75
11	.82	-.13	.69
12	.79	-.15	.65
13	-.36	.66	.57
14	.71	-.24	.56
Eigenvalue	6.77	2.35	9.13

^aRotated factor pattern matrix.^bh² = communality (sum of squared factor loadings) for variables.

Table 5.22

Combat Performance Prediction Scales Interrater Reliability Estimates

<u>Rating</u>	<u>1-Rater Reliability</u>	<u>n-Rater Reliability</u>
Supervisor Ratings	.42	.57
Peer Ratings	.24	.38
Combined Ratings	.31	.56

The mean and standard deviation of the Combat Scale scores are shown in Table 5.23 for the total second-tour sample, as well as for each MOS. The mean confidence rating across all raters was 5.33 (SD = 1.00) on a scale of 1 (not confident at all) to 7 (very confident).

Table 5.23

Combat Performance Prediction Scale Descriptive Statistics, by MOS

MOS	Supervisor			Peer			Combined		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
11B	110	68.9	12.9	66	65.3	12.1	112	68.2	12.1
13B	142	70.3	12.7	72	70.6	13.8	150	70.3	11.9
19E	32	72.7	13.3	21	76.1	11.8	33	73.4	12.0
19K	10	75.6	10.8	5	71.9	12.5	10	73.3	10.2
31C	75	69.5	13.4	29	68.1	10.8	84	69.0	12.8
63B	89	68.1	11.6	27	69.9	15.8	94	68.5	11.7
71L	41	71.1	13.4	5	83.0	13.0	41	71.5	13.7
88M	116	70.8	12.9	97	68.0	10.7	121	69.8	9.7
91A	68	73.2	11.9	21	75.8	10.0	70	73.3	11.2
95B	132	73.4	11.5	104	71.7	9.7	133	73.1	9.7
All MOS	815	70.8	12.6	447	70.0	12.0	848	70.2	11.5

Basic Scores for Second-Tour Hands-On Job Samples and Job Knowledge Tests

This section outlines the procedures used to formulate the basic criterion scores for the second-tour hands-on and job knowledge tests. The specific objectives were to (a) edit and prepare the data for analysis, and (b) combine the initial criterion scores into a shorter and more usable list of aggregated criterion scores, for use in constructing a model of job performance (see Chapter 6).

MOS Task-Specific Criterion Content

The procedures used to select second-tour tasks and to develop task tests are described in Campbell (1988), and are summarized only briefly here. Test content was generated by using all available information to define a population of tasks for each MOS. Sources of job- and task-analytic information included Soldier's Manuals (both MOS-specific and Common Task), Army Occupational Survey Program data on performance frequency, data on frequency and importance of supervisory tasks from administration of the Supervisory Responsibilities Questionnaire, and interviews with MOS incumbents. The resulting job domain included supervisory tasks, common tasks, and MOS-specific tasks. Tasks at lower skill levels were included in the domain because of the Army's policy that soldiers are responsible for such tasks; tasks at higher skill levels were included if there was evidence that soldiers in fact performed such tasks.

Judgments were obtained from subject matter experts on several task parameters, including performance difficulty, performance variability, and criticality. The task list for each MOS was clustered into functional areas,

and a second panel of SMEs selected representative samples of tasks stratified by functional category. These task samples were subjected to formal reviews by the proponent.

Multiple-choice job knowledge test items were constructed for about 30 tasks per MOS, and hands-on test situations were developed to test performance on 9-14 of these tasks per MOS. The tasks that were allocated to the hands-on component included, by design, both common and MOS-specific tasks, at skill level 1 as well as skill level 2 or higher, and from as many functional areas as was feasible for testing. The Multipurpose Arcade Combat Simulator (MACS) used for MOS 11B and 95B soldiers in LVI was also incorporated in the second-tour measures for soldiers in these MOS. The hands-on and job knowledge measures went through a series of pilot and field tests.

Supervisory tasks were not tested in either the hands-on or job knowledge component, but rather were covered by the Situational Judgment Test and the Supervisory Simulations (described later in this chapter).

Adjustment of Data

There were three sources of potential adjustments in item or task test scores from the CVII data collection: psychometrically marginal items, track differences, and missing data. Because of changes in equipment, changes in the proscribed steps in performance, and differences in performance under different conditions, not all test items were correct when the tests were administered--this despite rigorous tryouts and careful proponent agency review prior to the CVII data collection. Other items were simply far too easy or too difficult. In the former case, we were sometimes able to key one or more correct responses from among the alternatives offered, but some items had to be dropped. In the latter case, when items had pass rates of less than 15 percent or greater than 90 percent, or had negative discrimination indexes (Brogden-Clemans) within a task, the items were dropped. Table 5.24 shows the extent of the revisions or deletions.

Table 5.24

Revisions to Job Knowledge Components to Eliminate Marginal Items

	MOS								
	11B	13B	19E	31C	63B	71L	88M	91A/B	95B
Total Items	121	121	126	114	102	127	121	116	108
Items Dropped	3	0	2	2	4	3	1	9	11
Percent of Total	2.5	0.0	1.6	1.8	3.9	2.4	0.8	7.8	10.2
Items Revised	5	10	4	3	4	5	1	4	3
Percent of Total	4.1	8.3	3.2	2.6	3.9	3.9	0.8	3.4	2.8
Items Used	118	121	124	112	98	124	120	107	97

Different "tracks" within MOS were the second source of score adjustments or revisions. For hands-on tests, tracked versions were prepared as necessary to accommodate a requirement to use different types of equipment on the same task test. These tracks are, in essence, parallel versions of the same test. In some cases, equipment variations required only minor differences in performing a few hands-on steps, or the omission of a few steps. In other cases, although equipment and procedures were dissimilar across the tracks, the behavioral requirements were the same and the different tracks seemed to represent the jobs equally.

We examined the test results for evidence of task-determined level and dispersion differences across tracks by comparing differences in technical task scores to differences in basic task scores, which were not tracked. No anomalous differences were found, and consequently the data were adjusted by means of a linear transformation; specifically, the percentage of steps passed in a task was used as the task score. This transformation corrects for the variation in number of steps performed. Likewise, for the job knowledge tests, scores on tracked tests were adjusted for different numbers of items, by using percentage correct.

The third problem concerned missing data. For some tasks, difficulties in obtaining equipment for certain specialized tests at different sites precluded testing, either entirely or for segments of the task. If more than 15 percent of the tasks in the hands-on component were missing for a soldier, the soldier's data would not be used in analyses that included hands-on scores. For six of the MOS, the number of cases (soldiers) thus dropped from analysis was large (over 40%). Therefore, the decision was made to conservatively drop tasks, as necessary, for all soldiers but not to classify the entire hands-on record as missing. The objective was to maintain the sample size without losing the representativeness of the data with respect to the tasks selected for the MOS. The number of tasks thus dropped is shown in Table 5.25.

Table 5.25

Number of Tasks Dropped From Hands-On Component Due to Missing Data

	MOS								
	11B	13B	19E	31C	63B	71L	88M	91A/B	95B
Total Tasks	10	11	10	11	9	14	10	13	11
Tasks Dropped	1	1	0	1	3	2	0	1	0
Percent of Total	10.0	8.3	0.0	10.0	33.3	14.3	0.0	7.7	0.0
Tasks Used	9 ^a	10	10	10	6	12	10	12	11 ^a

^aIncludes MACS.

In the hands-on tests, data could also be missing for one of two other reasons: Either the scorer failed to observe a step, or the scorer failed to record the observation as either GO or NO-GO. In either event, the fact that the observation was missing was irrelevant to the soldier's performance. If few (i.e., fewer than 15%) of the steps on a given task were missing, then the remaining data were used as the soldier's score; computing percentage correct was used to adjust for the different number of steps. If, however, a larger number of steps was missing, the soldier's task score was set to missing for the task. Table 5.26 shows the extent of missing data on hands-on tests, after tasks were selectively dropped as shown in Table 5.25.

For the job knowledge tests, there were also two likely reasons for missing data: Either the soldier skipped an item (did not record a response on the answer sheet), or the soldier did not get to one or more items at the end of the test booklet. In the former case, we assumed that the soldier omitted the item because he or she did not know the answer; in those cases, the missing score was replaced by either a "guess" score (reciprocal of the number of alternatives) or by the actual proportion of soldiers passing the item, whichever was smaller. Although we were prepared to impute data for items not reached, the incidence was so small that items missing at the end of the test were treated the same as skipped items. Table 5.26 also shows the extent of missing data on the job knowledge tests.

Statistical Characteristics of Measures

The number of hands-on task tests and steps and the number of job knowledge task tests and items for each of the nine MOS are shown in Table 5.27. These are presented to support the display of statistical characteristics of the tests, as shown in Table 5.28 for each of the two measures for the nine MOS.

Because of the wide variation in numbers of steps per task, the hands-on task test scores are expressed as the percentage of task steps scored GO. The mean and standard deviation of the task scores (percentages) are also shown for each MOS, again expressed as a percentage to account for different numbers of tasks per MOS. In general, the overall means are somewhat higher than desirable, being over 70 percent for seven of the nine MOS. However, there was no evidence of extreme skew in the scores. The reliability index is a corrected split-half estimate, using halves that included both Basic and Technical tasks. The level of reliability was satisfactory, even considering the deliberate attempt to ensure heterogeneity in the tasks during task selection and hands-on test construction.

For the job knowledge tests, the results also include the range of task test means and standard deviations, with task test scores again expressed as percentage of items answered correctly. The number of items per task did not vary much (the average was just under four items per task, with a standard deviation of about 1.5); therefore, the overall mean and standard deviation across items (as opposed to across tasks) were computed. Again, the means are slightly above the desired 50 percent mark, but not extremely so. Reliabilities (odd-even items, balanced for Basic and Technical task items) were respectable, especially in view of the wide range of tasks covered by the tests.

Table 5.26

Extent of Missing Data on Hands-On and Job Knowledge Components

Component	MOS									Total
	11B	13B	19E	31C	63B	71L	88M	91A/B	95B	
<u>Hands-On</u>										
Tasks used	8 ^a	10	10	10	6	12	10	12	10 ^a	
Soldiers with no tasks missing	113	109	29	54	67	49	103	48	83	655
Percent of MOS soldiers	87.6	67.7	69.0	52.4	57.8	43.8	71.5	45.7	56.8	61.9
Soldiers with 25% or fewer tasks missing	11	30	0	34	33	56	26	41	52	283
Percent of MOS soldiers	8.5	18.6	0.0	33.0	28.4	50.0	18.1	39.0	35.6	26.7
Soldiers with over 25% of tasks missing	5	22 ^b	13 ^c	15	16	7	15	16	11	120
Percent of MOS soldiers	3.9	13.7	31.0	14.6	13.8	6.2	10.4	15.2	7.5	11.3
Total	129	161	42	103	116	112	144	105	146	1058
<u>Job Knowledge</u>										
Number of items	118	121	124	112	98	124	120	107	97	
Soldiers with no items missing	121	115	41	94	105	88	135	0	0	699
Percent of MOS soldiers	93.8	71.4	97.6	91.3	90.5	87.6	93.8	0.0	0.0	66.0
Soldiers with 10% or fewer items missing	4	19	1	2	5	16	1	101	138	286
Percent of MOS soldiers	3.1	11.8	2.4	1.9	4.3	14.3	0.7	96.2	94.5	27.0
Soldiers with over 10% of items missing	4	27	1	7	6	8	8	4	8	72
Percent of MOS soldiers	3.1	16.8	0.0	6.8	5.2	7.1	5.6	3.8	5.5	6.8
Total	129	161	42	103	116	112	144	105	146	1058

^aDoes not include MACS.

^bTwenty-two MOS 13B soldiers were not tested HO at Fort Campbell, where the howitzers were M102.

^cNine soldiers were MOS 19K who took a reduced HO, 10 tasks.

Table 5.27

Number of Hands-On Task Tests and Steps and Number of Job Knowledge Task Tests and Items for Nine MOS (Second Tour)

Component	MOS									Total
	11B	13B 109/110/198 ^a	19E	31C	63B	71L	88M	91A/B	95B	
<u>Hands-On</u>										
Task tests	8	10	10	10	6	12	10	12	10	
Steps	136	216/221/228	122	215	98	114+ NWPM ^b	191	210	258	
Range of steps per task test	7-47	6/8/8-53	4-20	2-70	7-43	2-43	6-47	7-43	7-53	
Mean steps per task test	17.0	22.1	12.2	21.5	16.3	10.4	19.1	17.5	25.8	17.8
SD of steps per task test	13.3	13.2	5.4	23.4	13.6	11.3	15.4	11.8	16.0	14.4
<u>Job Knowledge</u>										
Task tests	30	30	28	29	27	30	30	30	29	263
Items	118	121	124	112	98	124	120	107	97	1021
Range of items per task test	2-12	2-6	2-12	2-5	3-6	2-12	3-12	2-6	2-6	
Mean items per task test	3.93	4.03	4.43	3.86	3.63	4.13	4.00	3.57	3.34	3.88
SD of items per task test	1.74	1.19	1.83	.74	1.08	1.80	1.64	1.14	1.14	1.43

^aRepresents the three tracks for 13B testing.

^bNet words per minute on the task "Type straight copy."

Table 5.28

Statistical Characteristics of Hands-On and Job Knowledge Components for Each MOS

Component	MOS ^a								
	11B (129)	13B (161)	19E (42)	31C (103)	63B (116)	71L (112)	88M (144)	91A/B (105)	95B (146)
<u>Hands-On</u>									
Range of task test means (%)	46.4- 90.6	57.3- 87.1	35.1- 96.4	67.5- 86.1	61.2- 83.2	40.1- 79.9	52.5- 83.4	33.9- 86.3	67.4- 84.7
Range of task Test SDs (%)	13.9- 24.1	20.2- 39.0	7.2- 38.1	15.2 38.6	13.3- 35.0	19.0- 36.2	15.5- 32.4	12.1- 31.0	12.6- 29.3
Mean of task test means (%)	77.2	72.2	73.6	77.4	74.5	59.4	71.3	66.4	76.3
SD of task test means (%)	9.4	17.5	10.7	10.6	10.4	12.3	11.2	10.8	9.2
Reliability ^b	.51	.80	.57	.54	.20	.58	.65	.74	.51
<u>Job Knowledge Component</u>									
Range of task test means (%)	24.6- 90.1	42.1- 90.9	26.8- 90.5	34.2- 92.2	32.7- 87.1	44.6- 83.6	29.9- 90.6	31.9- 93.3	28.5- 91.2
Range of task Test SDs (%)	15.4- 31.1	18.5- 37.0	12.8- 36.5	15.7- 33.0	18.8- 37.7	18.3- 34.7	17.6- 32.7	16.8- 49.9	17.4- 48.8
Mean of items (%)	60.8	63.3	59.5	71.1	66.6	66.0	59.0	65.9	70.5
SD of items (%)	9.2	10.7	8.6	10.6	11.0	10.0	8.5	9.6	9.8
Reliability ^c	.83	.85	.83	.88	.87	.89	.80	.84	.83

^aNs are lower in some cells because of missing data.

^bSplit half reliability estimate, based on task test scores. Corrected to number of tasks.

^cSplit half reliability estimate, based on item scores. Corrected to number of items.

Construction of Basic Criterion Scores

Several analyses were performed to reduce the number of scores per individual (i.e., a reduction from 27-30 job knowledge task scores and 6-12 hands-on task scores). Principal component cluster analyses using task-level data (separately for each MOS and for job knowledge and hands-on tests) were largely uninterpretable for the knowledge tests. For the hands-on tests, the results yielded no consistent patterns across MOS.

Because no usable patterns emerged using task-level data, the tasks were clustered rationally on the basis of their content. The content considerations were based on the Functional Categories analyses which were developed for CVI (see Campbell, 1987, for a description), and which were also used to cluster the tasks for CVII task selection. The Functional Category rules developed for the CVI were modified slightly to accommodate differences between first-tour and second-tour job content. The Functional Category structure includes 10 across-MOS categories, plus one Technical Category per MOS (except 11B, where all tasks fit in the across-MOS categories).

Functional Category scores for job knowledge tests were formed as the proportion of items correct for all tasks assigned to the category. For hands-on tests, the category score was the mean of the hands-on task (proportion passed) scores. Principal component cluster analyses were performed for each MOS, using the Functional Category scores. Again, no consistent patterns emerged.

Continuing with a rational content-analytic approach (i.e., using task and category content rather than item response data to guide the reduction of the number of scores), we followed the procedures developed with the CVI data: Tasks were sorted into six higher-level groups referred to as Task Factors (Safety/Survival, Basic Techniques, Communication, Identify Targets, Vehicles, and Technical). Tasks were also combined into just two groups: Basic and Technical.

With two exceptions, the grouping schemes are hierarchical. That is, tasks (the lowest level) are placed in Functional Categories, the Functional Categories (level two) are aggregated to form the six Task Factors (level three), and Task Factors are then aggregated to form the two Task Constructs (level four). One exception involves the 11B Infantryman MOS, which has tasks in the across-MOS Functional Categories and in the non-Technical Task Factors; at the Task Construct level, however, the 11B tasks are all placed under the Technical heading, rather than under the Basic heading.

The second exception to the strictly hierarchical structure concerns the tasks involving vehicle maintenance and operation: The Task Factor for Vehicles is subsumed under the Basic Construct for three MOS where vehicle-related tasks are peripheral, but is subsumed under the Technical Construct for four MOS (13B Cannon Crewman, 19E Armor Crewman, 63B Light Wheel Vehicle Mechanic, and 88M Motor Transport Operator) where use of vehicles is central to the job. The hierarchical structure for the three grouping schemes is shown in Figure 5.3.

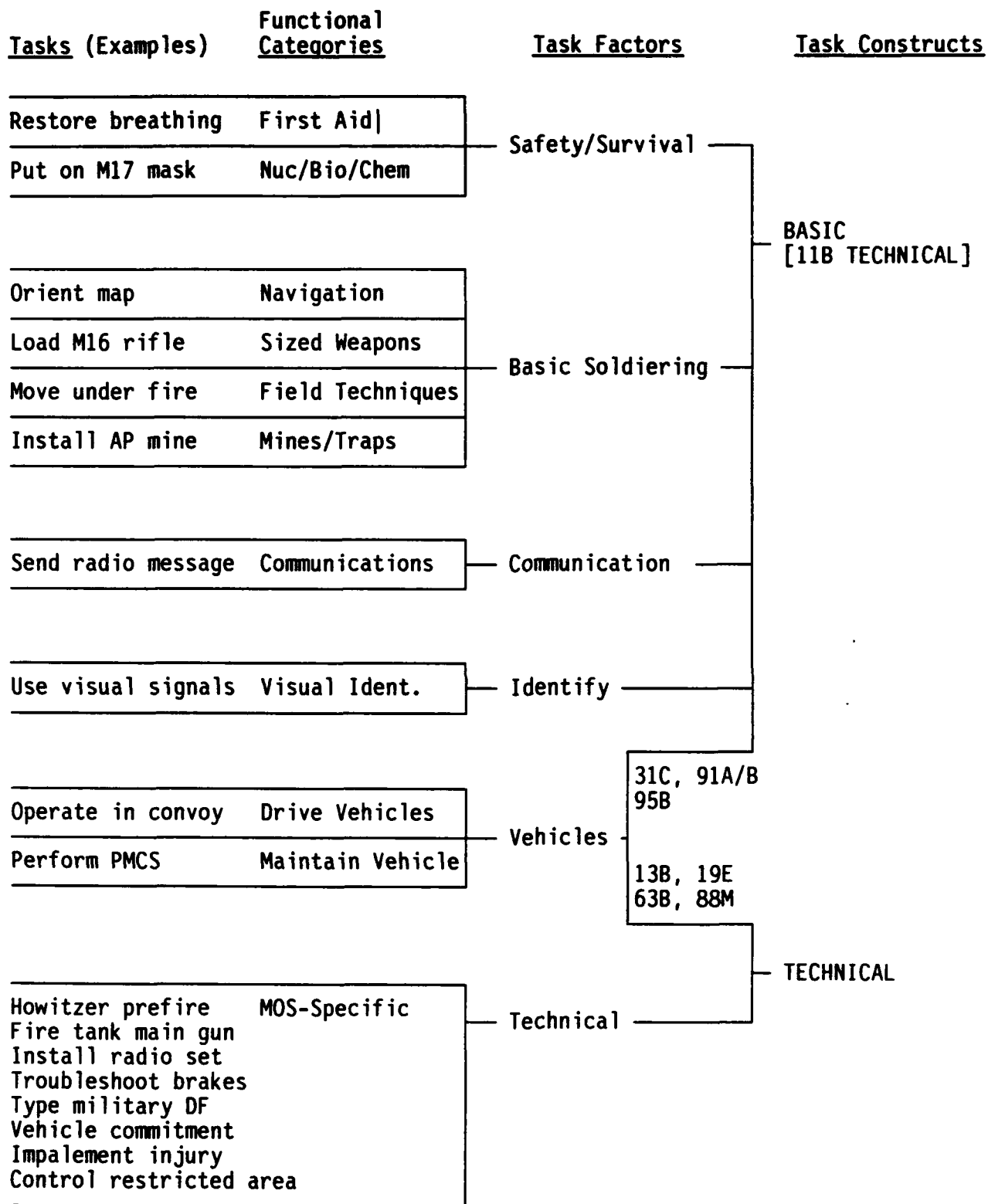


Figure 5.3. Hierarchical relationships among functional categories, task factors, and task constructs.

Not every MOS was represented by every Functional Category, or by every Task Factor. Table 5.29 summarizes the occurrence of Functional Categories and Task Factors for each MOS. With one exception (the 11B, as described above), every MOS has tasks in both the Basic and Technical Task Constructs. For job knowledge tests, the scores for each level are formed by computing the proportion of items passed across all tasks in the category. For hands-on tests, the score for each category is the mean of the task (proportion passed) scores.

Statistical characteristics of the hands-on and job knowledge test data for each level of aggregation (Functional Category, Task Factor, Task Construct) are shown in Tables 5.30, 5.31, and 5.32, respectively.

Several considerations argue for adoption of the highest level of aggregation--the two Task Constructs--in forming the basic criterion scores. First, data at that level are least affected by the differences in specific tasks tested across MOS. At the Functional Category level, the 11 categories by nine MOS should yield 99 cells; yet 44 of those cells are empty, where MOS did not have tasks assigned to the category. Similarly, at the Task Factors level, 18 of the 54 cells are empty. At the Task Construct level, only one of the 18 cells is not represented with data: the 11B Basic cell.

A second consideration concerns missing data. Within cells, small amounts of missing data can be replaced by imputed scores without doing violence to the variance-covariance matrix. However, if there are only a few scores comprising a cell, then even small amounts of missing data represent a sizable proportion of the total and even limited imputation may perturb the relationships.

The third consideration that favors adoption of the Task Construct scores is that it mirrors both the process and the solution adopted for the CVI performance model. For CVI, tasks were first grouped into functional categories on the basis of their similarity in task content. Next, scores were computed for each content category within each of the two test modes (hands-on and job knowledge). After category scores were computed, they were factor analyzed by means of principal components analysis. Separate factor analyses were executed for each type of measure within each job. The factors that emerged tended to be fairly similar across the nine different jobs and across the two test components.

As described above, the result of the first round of exploratory analyses in CVI was the set of six Task Factors. When scores constructed on these variables were subjected to another round of empirical factor analysis with other criterion variables (from various rating scales and administrative records), the hands-on and job knowledge tests each split between two higher order factors (common vs. MOS-specific) with the non-specific hands-on and job knowledge test Task Factor scores loading on the factor labeled General Soldiering Proficiency, and the MOS-specific hands-on and job knowledge test Task Factor scores loading on the factor labeled Core Technical Proficiency.

Table 5.29

MOS Task Representation in Functional Categories, Task Factors, and Task Constructs (Number of Job Knowledge Tests/Number of Hands-On Tests)

	MOS								
	11B	13B	19E	31C	63B	71L	88M	91A/B	95B
<u>Functional Categories</u>									
A. First Aid	3/1	2/1	1/1	3/1	2/1	3/1	3/1	3/1	1/0
B. Nuc./Bio./Chemical	3/1	3/0	4/1	4/1	3/0	4/1	5/1	3/0	3/1
C. Land Navigation	3/1	3/0	2/1	2/1	2/1	2/0	4/2	1/1	3/1
D. Sized Weapons	8/2	3/1	1/1	3/1	2/1	3/1	2/1	2/1	2/1
E. Field Techniques	8/1	3/0	2/0	2/0	1/1	3/1	5/0	4/0	5/1
F. Mines/Traps	2/1	2/1	1/0	0/0	0/0	0/0	2/0	0/0	1/1
G. Communication	2/1	1/0	2/1	2/2	1/0	2/2	1/1	2/2	1/0
H. Visual Identification	1/0	1/0	2/0	1/0	1/0	2/0	1/0	1/1	1/0
I. Drive Vehicles	0/0	1/0	0/0	1/0	0/0	0/0	4/2	0/0	0/0
J. Maintain Vehicles	0/0	1/0	1/1	0/0	2/0	0/0	1/1	1/1	1/1
X. MOS-Specific Tasks	0/0	10/7	12/4	11/4	13/2	11/6	2/1	13/5	11/4
<u>Task Factors (Comprising Functional Categories)</u>									
1. Safety/Survival (A + B)	6/2	5/1	5/2	7/2	5/1	7/2	8/2	6/1	4/1
2. Basic Soldiering (C + D + E + F)	21/5	11/2	6/2	7/2	5/3	8/2	13/3	7/2	11/4
3. Communication (G)	2/1	1/0	2/1	2/2	1/0	2/2	1/1	2/2	1/0
4. Identify (H)	1/0	1/0	2/0	1/0	1/0	2/0	1/0	1/1	1/0
5. Vehicles (I + J)	0/0	2/0	1/1	1/0	2/0	0/0	5/3	1/1	1/1
6. Technical (X)	0/0	10/7	12/4	11/4	13/2	11/6	2/1	13/5	11/4
<u>Task Constructs (Comprising Task Factors)</u>									
I. BASIC (1 + 2 + 3 + 4 + 5 ^a)	0/0	18/3	15/5	18/6	12/4	19/6	23/6	17/7	18/6
II. TECHNICAL (6 + 5 ^b)	30/8	12/7	13/5	11/4	15/2	11/6	7/4	13/5	11/4

^aMOS 31C, 91 A/B, 95B.

^bMOS 13B, 19E, 63B, 88M.

Table 5.30

Statistical Characteristics of Functional Categories for Hands-On and Job Knowledge Components for Nine MOS

Component	MOS									
	11B (129)	13B (161)	19E (42)	31C (103)	63B (116)	71L (112)	88M (144)	91A/B (105)	95B (146)	
<u>Hands-On</u>										
A. First Aid	Mean	79.6	80.0	91.5	78.3	75.6	74.0	77.2	56.6	--
	SD	18.2	20.7	17.1	18.3	23.8	31.3	22.5	25.1	--
B. Nuc./Bio./Chem.	Mean	89.4	--	44.6	80.8	--	79.9	80.1	--	67.5
	SD	15.5	--	18.9	22.4	--	23.7	19.8	--	29.3
C. Land Nav.	Mean	67.0	--	75.6	78.0	61.2	--	70.1	60.5	70.1
	SD	23.7	--	22.0	38.6	21.0	--	25.4	24.3	22.2
D. Sized Weapons	Mean	64.3	57.3	96.4	86.1	81.7	67.8	80.0	82.4	81.0
	SD	19.5	39.0	7.2	16.4	13.3	19.0	16.5	16.1	23.5
E. Field Tech.	Mean	83.6	--	--	--	83.2	76.2	--	--	80.3
	SD	19.3	--	--	--	16.5	19.9	--	--	17.7
F. Mines/Traps	Mean	90.6	83.9	--	--	--	--	--	--	73.6
	SD	13.9	20.2	--	--	--	--	--	--	22.1
G. Communication	Mean	79.4	--	35.1	77.2	--	31.8	66.6	39.5	--
	SD	18.7	--	38.1	24.5	--	32.7	23.4	30.5	--
H. Visual Ident.	Mean	--	--	--	--	--	--	--	79.2	--
	SD	--	--	--	--	--	--	--	24.5	--
I. Drive Vehicles	Mean	--	--	--	--	--	--	53.6	--	--
	SD	--	--	--	--	--	--	31.5	--	--
J. Mnt. Vehicles	Mean	--	--	72.1	--	--	--	83.4	72.5	82.2
	SD	--	--	11.6	--	--	--	17.9	20.7	17.1
X. MOS-Specific	Mean	--	71.8	82.5	73.9	72.9	52.8	74.2	72.1	77.3
	SD	--	22.1	8.8	16.4	25.9	14.9	15.5	12.1	9.6

(continued)

Table 5.30 (Continued)

Statistical Characteristics of Functional Categories for Hands-On and Job Knowledge Components for Nine MOS

		MOS								
		11B (129)	13B (161)	19E (42)	31C (103)	63B (116)	71L (112)	88M (144)	91A/B (105)	95B (146)
Component										
Job Knowledge										
A. First Aid	Mean	70.4	78.6	75.4	68.8	70.6	73.1	62.5	79.6	85.2
	SD	16.7	20.2	26.6	16.8	22.2	17.1	15.6	16.7	25.1
B. Nuc./Bio./Chem.	Mean	63.3	64.7	62.6	79.1	76.9	74.2	59.9	75.2	65.5
	SD	17.0	17.0	15.1	13.7	14.6	16.1	12.1	16.1	13.6
C. Land Nav.	Mean	76.3	60.4	67.9	72.4	49.3	64.1	60.2	51.7	65.0
	SD	20.7	19.0	20.0	26.3	24.2	25.6	15.3	30.9	20.6
D. Sized Weapons	Mean	47.5	62.9	73.4	77.2	74.8	64.8	73.8	80.4	48.2
	SD	9.4	16.2	20.5	16.5	18.7	17.4	18.2	18.5	14.9
E. Field Tech.	Mean	57.3	66.9	58.3	70.4	48.3	62.3	54.5	47.5	63.4
	SD	11.9	14.4	17.2	16.8	23.9	16.0	12.7	16.2	14.8
F. Mines/Traps	Mean	68.0	52.5	42.9	--	--	--	43.0	--	63.5
	SD	21.4	21.8	23.0	--	--	--	16.9	--	24.2
G. Communication	Mean	73.2	53.9	38.9	83.1	56.9	65.3	62.7	57.4	70.8
	SD	20.9	32.7	21.8	15.0	33.6	20.6	25.6	21.6	32.7
H. Visual Ident.	Mean	70.1	90.0	77.7	34.2	85.9	59.4	61.9	87.1	91.2
	SD	20.7	18.6	10.9	25.0	22.7	17.8	18.8	22.6	17.4
I. Drive Vehicles	Mean	--	43.9	--	80.5	--	--	57.9	--	--
	SD	--	23.1	--	23.7	--	--	14.5	--	--
J. Mnt. Vehicles	Mean	--	66.5	42.3	--	80.8	--	81.8	51.9	38.2
	SD	--	23.7	20.3	--	24.3	--	21.6	37.9	48.8
X. MOS-Specific	Mean	--	62.6	54.6	67.0	62.4	66.5	44.4	64.6	79.6
	SD	--	16.0	9.9	12.4	11.2	12.8	20.1	9.2	11.8

Table 5.31

**Statistical Characteristics of Task Factors for Hands-On and
Job Knowledge Components for Nine MOS**

		MOS								
		11B (129)	13B (161)	19E (42)	31C (103)	63B (116)	71L (112)	88M (144)	91A/B (105)	95B (146)
Component										
<u>Hands-On</u>										
1. Safety/ Survival	Mean	84.4	80.0	68.0	79.5	75.6	77.0	78.7	56.6	67.5
	SD	14.2	20.7	18.0	20.3	23.8	27.5	21.1	25.1	29.3
2. Basic Soldiering	Mean	73.8	70.6	86.0	82.1	75.4	72.1	73.4	71.4	76.25
	SD	10.9	29.6	14.6	27.5	16.9	19.4	22.5	20.2	21.38
3. Communication	Mean	79.4	--	35.1	77.2	--	51.8	66.6	39.5	--
	SD	18.7	--	38.1	24.5	--	32.7	23.4	30.5	--
4. Identify	Mean	--	--	--	--	--	--	--	79.2	--
	SD	--	--	--	--	--	--	--	24.5	--
5. Vehicles	Mean	--	--	72.1	--	--	--	63.6	72.5	82.2
	SD	--	--	11.6	--	--	--	27.0	20.7	17.1
6. Technical	Mean	--	71.8	82.5	73.9	72.9	52.8	74.2	72.1	77.3
	SD	--	22.1	8.8	16.7	25.9	14.9	15.5	12.1	9.6
<u>Job Knowledge</u>										
1. Safety/ Survival	Mean	66.5	68.6	64.4	74.9	74.4	73.8	60.8	77.3	68.6
	SD	12.8	16.3	14.8	12.8	14.3	13.1	10.8	12.7	12.6
2. Basic Soldiering	Mean	56.7	61.6	61.6	74.5	59.3	63.8	57.6	57.6	61.3
	SD	9.3	12.9	13.0	14.0	15.2	13.4	10.2	15.4	12.1
3. Communications	Mean	73.2	53.9	38.9	83.1	56.9	65.3	62.7	57.4	70.8
	SD	20.9	32.7	21.8	15.0	33.6	20.6	25.6	21.6	32.7
4. Identify	Mean	70.1	90.9	77.7	34.2	85.9	59.4	61.9	87.1	91.2
	SD	20.7	18.6	10.9	25.0	22.7	17.8	18.8	22.6	17.4
5. Vehicles	Mean	--	43.9	42.3	80.5	80.8	--	57.9	51.9	38.2
	SD	--	23.1	20.3	23.7	24.3	--	11.9	37.9	48.8
6. Technical	Mean	--	63.0	53.5	67.0	62.4	66.5	67.4	64.6	79.6
	SD	--	15.7	9.4	12.4	11.2	12.8	14.8	9.2	11.8

Table 5.32

**Statistical Characteristics of Task Constructs for Hands-On and
Job Knowledge Components for Nine MOS**

		MOS								
		11B (129)	13B (161)	19E (42)	31C (103)	63B (116)	71L (112)	88M (144)	91A/B (105)	95B (146)
<u>Hands-On</u>										
I. Basic	Mean	--	73.6	67.8	78.6	75.4	67.3	75.6	59.6	75.2
	SD	--	18.1	12.3	13.0	12.0	17.5	11.8	15.7	12.7
	Reliability ^a	--	.09	.34	.44	.45	.63	.57	.71	.47
II. Technical	Mean	76.9	70.6	77.0	74.0	72.7	53.1	62.1	71.9	76.3
	SD	9.8	22.2	14.4	16.1	25.0	14.8	18.4	12.5	10.6
	Reliability ^a	.51	.84	.43	.50	.43	.44	.32	.54	.34
<u>Job Knowledge</u>										
I. Basic	Mean	--	64.9	63.4	73.6	68.1	65.8	59.3	66.8	65.8
	SD	--	12.3	9.8	11.2	12.2	11.1	8.9	11.8	10.4
	Reliability ^b	--	.74	.76	.81	.84	.85	.74	.81	.78
II. Technical	Mean	60.8	61.1	53.8	66.5	65.5	66.5	57.9	65.5	79.6
	SD	9.2	14.3	9.2	12.7	12.1	12.8	11.9	9.4	11.8
	Reliability ^b	.83	.71	.63	.73	.80	.76	.66	.49	.78

^aReliability estimates for hands-on Basic and Technical scores are split-half estimates on tasks, corrected to test length.

^bReliability estimates for job knowledge Basic and Technical scores are split-half estimates on items, corrected to test length.

Consequently, for the CVII performance measures, four individual scores were constructed for each soldier: hands-on Basic Task score, hands-on Technical Task score, job knowledge Basic Task score, and job knowledge Technical Task score. Table 5.33 presents the means, standard deviations, reliability estimates, and correlations among the different scores, across the nine MOS.

The statistics shown in Table 5.33 present no surprises; reliability estimates (internal consistency) are higher for written job knowledge tests than for hands-on tests, due to the influence of the common method of testing. Similarly, the two construct scores for the written tests yield higher correlations than do the hands-on construct scores with each other and with the job knowledge test scores, again due to their written mode of testing. Although the correlations (uncorrected) are not outstandingly high, the within-construct/across-method correlations (i.e., the two Basic scores and the two Technical scores, between hands-on and job knowledge) are higher than the across-construct correlation for hands-on tests. These findings lend support to the claims for two measured constructs of performance in the test results.

Summary

Initial analyses of the second-tour data were conducted for the job knowledge tests and hands-on tests. The objective was to reduce large sets of task, item, and scale scores to the set of basic criterion scores that would be used with the other criterion data to develop the final criterion factor scores. Our analyses were directed at results at the task test level, at functional categories of tasks, and at higher aggregations of task scores.

Factor analyses of the tasks and functional category data were severely hampered by the small sample sizes. Rational content analyses were performed that mimicked the CVI factor analyses, resulting in a structure comprised of two scores for each of the two test modes: hands-on Basic Task scores and Technical Task scores, and job knowledge Basic Task scores and Technical Task scores. Scores on these composites of tasks form the basic criterion scores for further modeling of second-tour performance.

Basic Scores for Administrative Measures

The Personnel File Form (PFF) is an instrument used to gather self-report administrative information. The first-tour version of the PFF gathered information on the following items:

- (1) Awards (e.g., Army Achievement Medal)
- (2) Memoranda and Certificates of Appreciation, Commendation, and Achievement
- (3) Negative personnel actions (i.e., Articles 15 and Flag Actions)
- (4) Army Physical Fitness Test score

Table 5.33

Statistical Characteristics of Hands-On and Job Knowledge Component Basic Task Scores and Technical Task Scores Across Nine MOS^a

Component	Mean	SD	Reliability
<u>Hands-On</u>			
Basic Task Score	72.9	15.2	.47
Technical Task Score	69.2	18.7	.48
<u>Job Knowledge</u>			
Basic Task Score	65.2	12.4	.80
Technical Task Score	65.5	13.8	.70

Mean Correlations

Component	<u>Job Knowledge Component</u>		<u>Hands-on Component</u>	
	Basic	Technical	Basic	Technical
<u>Job Knowledge</u>				
Basic Tasks	1.00			
Technical Tasks	.55	1.00		
<u>Hands-On</u>				
Basic Tasks	.34	.24	1.00	
Technical Tasks	.24	.30	.20	1.00

Note. Based on 42-161 soldiers per MOS; total number of soldiers is 1058.

^aMeans, standard deviations, reliabilities, and correlations shown are averages across the nine MOS. Correlations are not corrected for unreliability.

- (5) M16 qualification category
- (6) Skill Qualification test score

Second-Tour Personnel File Form

In addition to the items explored in the first-tour data, the second-tour version of the PFF gathered information on military training courses, civilian education, and promotions. Given that this was essentially a field test version of the form, each type of information on the PFF was examined to determine its suitability for scoring. Primary considerations were the distributional characteristics, interpretability, and accuracy of the self-report information.

Positive Recognition Items. PFF items that reflected recognition of exceptional job performance included those in the awards section and the section covering memoranda and certificates of appreciation, commendation, and achievement. The awards section consisted of a checklist of awards, along with three blank spaces in which respondents could write in awards not listed on the form. Review of the write-in entries indicated that most were awards for which the soldiers should not be given credit, such as those which soldiers would get for simply "being there." For example, the Army Overseas Service Ribbon is given to all soldiers who serve overseas, regardless of how they perform while they are there. Other write-in entries (e.g., Recruiting Badge), however, were considered suitable for scoring and were added to the scoring algorithm.

The NCO promotion board system gives credit for awards received by soldiers. In this process, awards are weighted based on their importance to the Army. The awards subscore for the PFF was computed by weighting the awards using the promotion board weights and summing across all scored awards. This increased the variability of the awards score (compared to computing an unweighted sum) and appeared to more accurately reflect the relative job performance of the respondents.

Other evidence of positive job performance is the receipt of memoranda and certificates which cite specific situations in which the soldier has performed well. On the PFF, respondents indicated how many memoranda and certificates they had received while at different paygrades. In an effort to better distinguish between first- and second-tour performance, only memoranda and certificates received while in grades E-4 and above were scored. This subscore was a simple sum of the number of memoranda and certificates that had been received while the soldier was an E-4 or above.

The awards and memoranda/certificates subscores were correlated .29 with each other ($n = 927$). Considering this degree of covariance as well as their conceptual similarity, the awards subscore and the memoranda and certificates subscore were combined to create a single criterion score called "Awards".

Disciplinary Actions. An Article 15 is a disciplinary personnel action and a Flag Action is the suspension of a favorable personnel action. Both are considered to indicate poor soldier performance. As with the memoranda and certificates, PFF respondents were asked to indicate how many Articles 15 and Flag Actions they had received while at different paygrades. The PFF "Article

15/Flag Action" score was computed by summing the number of Articles 15 and Flag Actions the soldier received while in paygrades E-4 and above.

Army Physical Fitness Test (APFT). All soldiers are required to take a test of physical readiness annually. This test includes push-ups, sit-ups, and a 2-mile run. The APFT score is derived from a conversion table which adjusts the measurements based on the age and gender of the soldier. The self-report APFT score from the PFF is referred to as the "Physical Readiness" score.

M16 Qualification. All soldiers are periodically tested on their M16 (or M1911A1) marksmanship. On the basis of this test, they are categorized as marksmen, sharpshooters, or experts. The PFF asked respondents to report their most recent marksmanship classification. This comprised the PFF "M16 Qualification" score.

SQT Score. Skill Qualification Tests have normally been given once a year to provide an operational assessment of each soldier's job proficiency. Typically, the tests covered both MOS-specific and Army-wide tasks and were composed of multiple-choice items. Although the correlation between the self-report SQT from the PFF and the SQT score available from the Enlisted Master File was reasonable ($r = .80$, $n = 904$), it was decided that the SQT score would not be used as a PFF score, primarily because it was an operational performance measure, and thus a possible independent criterion.

Military Courses and Civilian Education. The second-tour PFF included a checklist of military training courses, such as preliminary and advanced NCO training and Ranger School. A military training score was computed by summing the checked items on the list.

Respondents were also asked to indicate how many semester hours they had of business school, trade school, and college courses. This section of the PFF showed a significant amount of missing data and the accuracy of the soldiers' responses was considered to be questionable for several reasons. For example, even if a soldier was able to remember the number of hours he or she had earned, it is unlikely that soldiers who attended schools on a quarter system would have been able to accurately convert these hours into semester hours. Since there was no way to assess the accuracy of the information, it was not scored.

Promotions. PFF respondents were asked to indicate whether they had ever been recommended for a promotion in the secondary zone. This would mean that they had been recommended by their commander for a promotion before they had the requisite time in grade. Soldiers to whom this had happened were asked to indicate at what grade and time in service they had been given this recommendation. A second measure of promotion rate was computed for each soldier based on information in computerized personnel records (i.e., the Enlisted Master File). This is a grade deviation score in which each individual's paygrade is adjusted to the mean of those who share his or her time in service. These two measures were combined to form the PFF "Promotion Rate" score.

To be promoted to an E-5 or higher rank, soldiers must go through a promotion board review process, which includes an appearance before a promotion board, and a paper application in the form of a promotion board

worksheet. Soldiers receive administrative points based on the contents of the worksheet and promotion board points based on their board appearance. Although self-reports of both sets of points were collected on the PFF, none of this information was scored. In addition to concerns regarding the accuracy of the self-report information (and no adequate way to verify the accuracy), the administrative points showed relatively little variability. The maximum number of points was 200 and the median score was 190 (mean = 187, SD = 13.08, N = 719). It was decided that these points were too redundant with other PFF scores because they are computed using much of the same information (i.e., awards, memoranda and certificates, Physical Fitness Test score, SQT score, military training, and M16 qualification).

Summary of Personnel File Form Scores for CVII

In summary, six scores were derived from the second-tour PFF: Awards, Articles 15/Flag Actions, Physical Readiness, M16 Qualification, Military Training, and Promotion Rate. In the Concurrent Validation, the first-tour PFF had been scored so that it yielded variations on five of these scores (i.e., Awards, Articles 15/Flag Actions, Physical Readiness, M16 Qualification, and Promotion Rate). Table 5.34 lists the means and standard deviations of each of the second-tour scores and Table 5.35 shows the intercorrelations among them.

Table 5.34

Second-Tour Personnel File Form Score Means and Standard Deviations

	N	Mean	SD
Awards	927	10.53	5.63
Article 15/Flag Action	929	.42	.87
Physical Readiness	997	250.16	30.66
M16 Qualification	1035	2.52	.67
Military Training	1060	1.35	1.03
Promotion Rate	928	100.15	8.11

Table 5.35**Second-Tour Personnel File Form Score Intercorrelations**

	Awards	Art 15	Physical	M16	Train	Promo
Awards	-					
Article 15/Flag Action	-.08	-				
Physical Readiness	.13	-.11	-			
M16 Qualification	.14	-.03	.11	-		
Military Training	.31	-.16	.19	.19	-	
Promotion Rate	.31	-.19	.14	.14	.39	-

Note. Ns range from 817 to 1,035; all correlations except -.03 are significant at $p < .01$.

Personnel File Form Revisions for LVII

Based on the scoring decisions described here, a number of changes will be made to the PFF before it is administered to the second-tour longitudinal sample in 1991. The changes will be aimed at eliminating information that is unlikely to become part of the second-tour performance model, clarifying items so that missing data are less of a problem, and improving the usefulness of information by altering items regarding training, promotions, and perhaps civilian education.

The Situational Judgment Test

This section reports the analyses of Situational Judgment Test (SJT) data for the CVII sample. This multiple-choice paper-and-pencil test was developed as a criterion measure of supervisory skill for noncommissioned officers. The collection and analyses of the SJT data from the CVII sample had three major objectives: (a) to examine and evaluate the psychometric properties of this instrument; (b) to develop one or more SJT scores to be used in the modeling of second-tour performance; and (c) to conduct preliminary investigations into the construct validity of the SJT as a criterion measure of supervisory job knowledge.

Situational Test Content

Situational judgment tests have been developed by other researchers as predictors of job performance, especially for management and supervisory

positions (e.g., Motowidlo, Carter, & Dunnette, 1989; Mowry, 1964; Rosen, 1961; Tenopir, 1969). In contrast, the SJT was developed as a criterion measure. Previous researchers have found that scores on written simulations differentiate between groups with differing levels of experience or training (e.g., Alderman, Evans, & Wilder, 1981; McGuire & Babbott, 1976) and are often related to other measures of professional knowledge or performance (see Smith, 1983 for a review).

The test used in the Career Force project, the SJT, consists of 35 items. For each item, soldiers are asked to read a description of a difficult supervisory situation, examine three to five possible responses to the situation, then select the most and the least effective response alternatives. The following example is representative of the kind of items that make up the SJT (this is not an actual SJT item).

You are a squad leader on a field exercise, and your squad is ready to bed down for the night. The tent has not been put up yet, and nobody in the squad wants to put up the tent. They all know that it would be the best place to sleep because it may rain, but they are tired and just want to go to bed. What should you do?

- a. Tell them the first four men to volunteer to put up the tent will get light duty tomorrow.
 - b. Make the squad sleep without tents.
 - c. Tell them that they will all work together and put up the tent.
 - d. Explain that you are sympathetic with their fatigue, but the tent must be put up before they bed down.
-

These items are intended to evaluate the effectiveness of NCO judgments about what to do in difficult supervisory situations. Thus, the SJT is similar to a job knowledge test for supervisory job content.

As reported previously (Campbell, 1989), and summarized in the introduction to this chapter, development of the SJT involved asking groups of soldiers similar to the target NCOs (i.e., at the E-4 and E-5 level) to describe a large number of difficult but realistic situations that Army first-line supervisors face on their jobs. Once a large number of these situations had been generated, a wide variety of possible actions (i.e., response alternatives) for each situation were gathered, and ratings of the effectiveness of each of these actions were collected from both experts (senior NCOs) and the target group (E-4 and E-5 NCOs in beginning supervisory positions). These effectiveness ratings were used to select situations and response alternatives to be included in the SJT.

The sample of experts was a group of 90 senior NCOs who were students and instructors at the Sergeants Major Academy. These NCOs were some of the highest ranking enlisted soldiers in the Army (rank of E-8 to E-9), and they

all had extensive experience as supervisors in the Army, with an average of over 15 years in supervisory positions. Thus, they were in an ideal position to provide information about the actual effectiveness of the various SJT response alternatives. For each situation, these NCOs rated the effectiveness of each response alternative on a 7-point scale (1 = least and 7 = most effective). Because about 180 situations (i.e., items) were still being considered at the time these data were collected, each NCO rated the response alternatives for only a subset of the items that are currently included in the SJT. Thus, about 25 expert judgments were available for each of the SJT items. Again, 35 items were selected for the final form.

The SJT Sample

The SJT was administered to a total of 1049 soldiers from the CVII sample. Eleven percent of these soldiers were women, and the racial breakdown was as follows: 56 percent white, 33 percent black, and 6 percent Hispanic (the remainder reported "other"). For each SJT item, these soldiers were told to mark an "M" next to the response alternative they thought was the most effective and an "L" next to the response alternative they thought was the least effective.

Data Analysis Procedures

These data were first screened for invalid (e.g., two M scores or two L scores) and incomplete data. Next, frequency counts were conducted of the number and percentage of respondents choosing each SJT response alternative, to determine whether there was variability in the answers chosen by respondents in this sample. Because of the multiple-choice format of the SJT items, it was conceivable the correct answer was obvious (i.e., the test is too easy). If this were the case, it would be impossible for SJT scores to discriminate among these soldiers.

Development of Scoring Procedures. Several different procedures for scoring the SJT were explored. The most straightforward was a simple "number correct" score. For each item, the response alternative that the experts had given the highest mean effectiveness rating was designated the "correct" answer. Respondents were scored on the number of items for which they indicated that the "correct" response alternative was the most effective.

The second scoring procedure involved weighting each response alternative by the mean effectiveness rating given to that alternative by the expert group. This gives respondents more credit for choosing "wrong" answers that are relatively effective than for choosing wrong answers that are very ineffective. These item-level effectiveness scores were then averaged to obtain an overall effectiveness score for each soldier. Averaging these item-level scores instead of simply summing them placed respondents' scores on the same 1 to 7 effectiveness scale as the experts' ratings and ensured that respondents were not penalized for missing data.

Scoring procedures based on respondents' choices for the least effective response to each situation were also evaluated. The ability to identify the least effective response alternatives might be seen as an indication of respondents' ability to avoid these very ineffective responses--in effect,

to avoid "messing up." As with the choices for the most effective response, a simple number correct score was computed: the number of times each respondent correctly identified the response alternative that the experts rated the least effective. This score will be referred to as the L-Correct score, and the score based on choices for the most effective response as the M-Correct score.

Another score was computed by weighting respondents' choices for the least effective response alternative by the mean effectiveness rating for that response, and then averaging these item-level scores to obtain an overall effectiveness score based on choices for the least effective response. This score will be referred to as L-Effectiveness, and the parallel score based on choices for the most effective responses as M-Effectiveness.

Finally, a scoring procedure combining the choices for the most and the least effective response alternative into one overall score was investigated. For each item, the mean effectiveness of the response each soldier chose as the least effective was subtracted from the mean effectiveness of the response chosen as the most effective. Because it is actually better to indicate which less effective response alternatives are the least effective, this score can be seen as a composite of the two effectiveness scores (i.e., subtracting a negative number from a positive number is the same as adding the absolute values of the two numbers). These item-level scores were then averaged together for each soldier to generate yet another score, to be referred to as M-L Effectiveness.

Descriptive statistics and two estimates of internal consistency (KR-20 and split-half) were computed for each of these five scoring procedures. To compute split-half reliabilities, two researchers each divided the SJT into two "parallel" halves according to the content of the item stems and response alternatives. Correlations were then computed between the two parallel halves identified by each of these researchers and corrected for test length using the Spearman-Brown formula. Intercorrelations were also computed among the five scores generated by the five different scoring procedures. Finally, item analyses were conducted for each of the scoring procedures. These item analyses included the item-total correlations for all of the scoring procedures and also the proportion of the sample answering each item correctly for the M- and L-Correct scoring procedures.

Subgroup Analyses. Subgroup analyses were also conducted for the three most promising scoring procedures (M-Effectiveness, L-Effectiveness, and M-L Effectiveness). Descriptive statistics for each of these scoring procedures were computed separately for males and females and for several different racial subgroups. Descriptive statistics were also computed separately for soldiers from combat and non-combat MOS, and for soldiers from each of the nine MOS included in the CVII sample.

These latter analyses will provide information concerning whether the SJT is an equally appropriate measure of supervision for all nine MOS. Some of the participants in the SJT development workshops indicated that supervision in combat MOS is somewhat different than supervision in non-combat MOS. For example, some of them reported that supervisors in combat MOS are expected to take a stricter approach to subordinate misconduct. If the "correct" answer to SJT items varies by MOS, this may be reflected in differences in the mean scores of soldiers from different MOS.

Preliminary Investigations of Construct Validity. One of the objectives of the present research was to obtain preliminary evidence concerning whether this test does in fact measure supervisory job knowledge. Three different approaches were taken to obtaining information about construct validity.

First, because the SJT response alternatives describe a variety of different types of supervisory behavior (e.g., counseling, disciplining, planning and organizing), it is possible that the SJT is not unidimensional but actually measures several relatively distinct subconstructs. Consequently, the dimensionality of the SJT was investigated. If distinct subconstructs were identified, they could provide the basis for the development of SJT dimension scores. Item-level scores for the 35 SJT items were factor analyzed, but even within a single item the response alternatives appear to represent a variety of different supervisory behaviors. Therefore, the dimensionality of the SJT was also explored at the response alternative level. Respondents were given a score for each response alternative indicating whether or not they chose that alternative (e.g., 0 or 1) and the response alternative scores were then factor analyzed as well.

The second approach was to obtain ratings of the supervisory behaviors represented by the SJT response alternatives to determine the extent to which the SJT taps the various aspects of supervision identified in the second-tour job analyses (Campbell, 1989). This was done by first identifying a set of dimensions of supervisory behavior relevant to the SJT. Four researchers were asked to independently content analyze the SJT response alternatives, develop a category system based on the supervisory behaviors involved, and categorize the SJT response alternatives into these categories or dimensions. These four category systems were then rationally collapsed into a single system with 10 dimensions. Table 5.36 shows these 10 dimensions and provides a definition of each.

Nine researchers were then asked to rate the extent to which each SJT response alternative taps each of these 10 dimensions. Because some alternatives appeared to involve more than one dimension, each alternative was assigned 10 points, and raters were asked to divide these points among the dimensions according to the content of the response alternative. This also ensured that some responses would not be over-represented in the final category system. Finally, these raters were told to assign the points from response alternatives that didn't fit into any of these categories to a miscellaneous category.

The interrater reliability of these ratings was computed for each of the 10 dimensions, and the mean ratings for each response alternative were used to calculate the extent to which the SJT taps each of the 10 dimensions. In addition, to determine whether certain supervisory behaviors tend to be more effective than others, the average effectiveness of each of the 10 dimensions was computed, using these ratings in combination with the effectiveness ratings from the USASMA experts.

One final type of information used to assess the construct validity of the SJT was the extent to which the knowledges assessed by the SJT are learned on the job. Because the SJT is intended to be a criterion measure of job knowledge, soldiers who have more experience or training would be expected, on average, to obtain higher scores than soldiers with less experience or training. Self-report information was collected from the CVII soldiers

Table 5.36

Situational Judgment Test: Supervisory Behavior Dimension Definitions

-
1. *Referring.* Refer subordinates to a counseling or help program (e.g., financial counseling, a dietitian, the education center, formal counseling) in response to personal or performance problems.
 2. *Interacting assertively with superiors.* Work assertively with individuals at a higher level in the chain of command (e.g., to stick up for subordinates' rights, obtain appropriate rewards and punishments for subordinates, or solve subordinates' problems).
 3. *Counseling.* Conduct formal or informal counseling with subordinates concerning performance or personal problems. This includes disciplinary counseling as well as counseling meant to encourage subordinates, help them solve problems, etc.
 4. *Encouraging.* Provide encouragement to subordinates by acknowledging or rewarding good performance or exemplary behavior, also by providing encouragement and support in response to their problems.
 5. *Disciplining.* Discourage inappropriate behaviors or inadequate performance by taking disciplinary actions (e.g., Articles 15, formal counseling statements, additional duty), by warning that disciplinary action may be taken in the future, or by reporting the problem to superiors.
 6. *Gathering information/monitoring.* Gather the information necessary to strategically assign tasks or to take action in response to problems (e.g., poor performance). Monitor subordinates' performance or other behaviors.
 7. *Reasoning with soldiers.* Ensure that subordinates perform assigned tasks and duties by reasoning with them (e.g., explaining why the work must be done, providing an incentive, or otherwise persuading them).
 8. *Giving orders.* Give soldiers direct orders (e.g., orders to perform tasks, activities, or missions).
 9. *Assigning tasks.* Strategically assign tasks in a manner that will best accomplish the mission, address subordinate problems (e.g., performance, personal, or interpersonal problems), or provide developmental opportunities.
 10. *Communicating with subordinates.* Provide subordinates with needed information or advice; keep subordinates informed. This includes communicating specific performance expectations, clarifying tasks or missions, or telling subordinates about opportunities that are available to them.
-

concerning how long they had been in a supervisory position, how regularly they were required to supervise other soldiers, and whether they had attended any supervisory training. The mean SJT scores of soldiers with differing levels of experience and training were examined, and the correlations of SJT scores with amount of supervisory experience were also examined.

Results of SJT Analyses

Data Screening. While the majority of the 1049 inventories were completed correctly, some data editing was required. SJT scores were then computed only for soldiers who had valid item-level data for at least 90 percent of the items. A total of 1,025 soldiers had valid "M" responses for at least 90 percent of the items, 1,007 had valid "L" responses for at least 90 percent of the items, and 1,007 had both valid "M" and valid "L" responses for at least 90 percent of the SJT items.

Item-Level Frequencies. The SJT item-level responses from the CVII sample were distributed quite well across the response alternatives for each item. For example, the percentage of respondents choosing the most popular response alternative for each item as the most effective ranged from 32 to 74, with a median of 46 percent. The correct responses to SJT items were not at all obvious to the soldiers in this sample.

Descriptive Statistics for the Five Scoring Procedures. The mean score for each of the five scoring procedures is presented in Table 5.37. The maximum possible for M-Correct is 35 (i.e., all 35 items answered correctly), but the maximum score obtained in this sample was only 27, and the mean was 16.25. The mean number of least effective response alternatives correctly identified by this group was only 14.86. Clearly the SJT was difficult for this group of soldiers. The fact that the mean of L-Correct is lower than M-Correct implies that identifying the least effective response alternative was more difficult than identifying the most effective alternative. However, it should be noted that SJT items and response alternatives were selected from the total pool of items and response alternatives based on respondents' choices for the most effective response alternative. As a result, the least effective response alternatives often have L-Effectiveness levels that are similar to one or more other response alternatives. The L-Effectiveness score is probably a more appropriate score, because respondents are given credit for identifying these other relatively ineffective response alternatives.

Table 5.37 also presents the standard deviation and the minimum and maximum scores obtained for each scoring procedure. All procedures generate reasonable variability. The table also shows that the internal consistency reliabilities for all scoring procedures are quite high; split-half reliabilities are at about the same level as the KR-20 reliability estimates. If the SJT was a multidimensional test and if the two halves used to compute these reliabilities were truly parallel, we would expect these split-half reliabilities to be higher than the KR-20 reliability estimates. The fact that they are nearly equal may indicate that the parallel halves identified by researchers were not truly parallel, or it may indicate that the SJT measures a fairly unidimensional construct. The most reliable score is M-L Effectiveness, probably because this score contains more information than the other scores (i.e., choices for both the most and least effective response).

Table 5.37

Situational Judgment Test: Means and Internal Reliabilities

Scoring Procedure	N	Max. Score	Min. Score	Mean	SD	KR-20 Internal Consistency	Split-Half ^a Reliability	
							A	B
M-Correct (no. of correct "Most" responses)	1025	27	3	16.52	4.29	.60	.60	.60
M-Effectiveness (mean eff. scale value for "Most" responses)	1025	5.65	3.66	4.91	.34	.68	.68	.68
L-Correct (no. of correct "Least" responses)	1007	25	2	14.86	3.86	.57	.53	.48
L-Effectiveness (mean eff. scale value for "Least" responses)	1007	2.90 ^b	4.84 ^b	3.54 ^b	.31	.68	.70	.67
M-L Effectiveness (mean of eff. scale value for "Most" responses minus eff. scale value for "Least" responses)	1007	2.57	-.77	1.36	.61	.75	.77	.75

^aSplit-half reliability estimates are the correlations between two "parallel" halves corrected for test length using Spearman-Brown.

^bLow scores are "better"; mean effectiveness scale values for L-responses should be low.

The intercorrelations among scores obtained using the five different scoring procedures are presented in Table 5.38. They range from moderate to high. Correlations between scores that are based on the same set of responses (e.g., M-Correct with M-Effectiveness) are higher than correlations between scores that are based on different sets of responses (e.g., M-Correct with L-Correct). The correlation between L-Effectiveness and the other indexes is negative because lower scores on this index are actually better. The high (negative) correlation between M-Effectiveness and L-Effectiveness seems to indicate that these two scores measure similar or related constructs.

Table 5.38

Situational Judgment Test: Score Intercorrelations for the Five Scoring Procedures

	M-Eff.	L-Correct	L-Eff.	M - L Eff.
M-Correct	.94	.52	-.64	.86
M-Effectiveness	--	.59	-.70	.93
L-Correct	--	--	-.86	.78
L-Effectiveness	--	--	--	-.92

Note. Sample sizes range from 1007 to 1025.

The median and range of the 35 item-total correlations obtained using each of the scoring procedures are shown in Table 5.39. These correlations are reasonably high, although there is considerable variability across items. As would be expected, the scoring procedures that yield more internally consistent scores also have, on average, higher item-total correlations. For three SJT items, the item-total correlations were extremely low for at least one scoring procedure (ranging from .00 to .09). For each of the five scoring procedures, scores were recomputed excluding these suspect items; internal consistency reliabilities increased by only about .01, so all items were retained for the remainder of the analyses. For the M- and L-Correct scoring procedures, the proportion of the sample answering each item correctly ranged from less than 25 percent to 74 percent.

The M-Correct and L-Correct scores have somewhat less desirable psychometric characteristics than the other three procedures. In addition, these two scores contain information that is very similar to the information provided by the M-Effectiveness and L-Effectiveness scores respectively, because they are based on the same sets of responses. Results reported for the remainder of the analyses will not include the M- and L-Correct scores.

Table 5.39**Situational Judgment Test: Summary of Item Analysis Results**

Scoring Procedure	Item-Total Correlations		Proportion Answering Items Correctly	
	Range	Median	Range	Median
M-Correct	.03 to .47	.25	.24 to .74	.44
M-Effectiveness	.06 to .51	.29	--	--
L-Correct	.05 to .47	.21	.15 to .71	.40
L-Effectiveness	.00 to .54	.27	--	--
M-L Effectiveness	.05 to .52	.33	--	--

Subgroup Analyses. For the three remaining scoring procedures, Table 5.40 shows that females tend to score, on average, about a third of a standard deviation higher than males. Analysis of variance revealed that these differences are significant, but do not account for a great deal of the variance in SJT scores. In addition, blacks scored about a third of a standard deviation lower than whites and Hispanics. Analysis of variance showed that these differences among racial groups are also significant but do not account for much variance.

The mean SJT scores for soldiers in combat and non-combat MOS are shown in Table 5.41. For the combat MOS (11B, 13B, and 19E/K) mean SJT scores are about a quarter of a standard deviation lower than for the other five MOS. These differences are significant, but account for very little variance.

Table 5.41 also shows the mean SJT scores for each of the nine individual MOS. The MOS with the highest mean scores are 95B and 19E/K, and the MOS with the lowest mean scores include 13B and 88M. Analysis of variance showed that these differences are significant, and they account for more variance than the combat/non-combat differences. This result might be explained in part by differences in general cognitive ability. Different MOS have different selection standards and Table 5.41 shows the mean Armed Forces Qualification Test (AFQT) scores for the various MOS. It also shows the rank order correlations, across MOS, between the mean AFQT score and mean SJT scores. These suggest that differences in cognitive ability can account for at least some of the differences in SJT scores across MOS.

Table 5.40

Situational Judgment Test: Scores for Demographic Subgroups

Subgroup	N	M-Effectiveness			L-Effectiveness			M-L Effectiveness		
		Mean	SD	R ²	Mean	SD	R ²	Mean	SD	R ²
Male	873-867	4.89	.34		3.55	.31		1.35	.60	
Female	105-109	5.04	.30	.02 *	3.48	.34	.004	1.55	.58	.01 *
Black	316-324	4.81	.34		3.63	.33		1.19	.61	
Hispanic	61-63	4.97	.32		3.54	.31		1.43	.55	
White	550-557	4.96	.34		3.48	.29		1.48	.58	
Other ^a	51-52	4.90	.31	.04 *	3.62	.33	.05 *	1.29	.58	.05 *

*Significant at the .01 level.

^aThe amount of variance accounted for by race was estimated using only blacks, Hispanics, and whites.

Table 5.41

Situational Judgment Test: Combat/Non-Combat and MOS Differences in Scores

MOS	N	M-Effectiveness ^a			L-Effectiveness ^b			M-L Effectiveness ^c			AFQT Mean
		Mean	SD	R ²	Mean	SD	R ²	Mean	SD	R ²	
Combat MOS ^d	278-309	4.84	.34		3.58	.32		1.27	.61		49.55
Non-Combat MOS ^e	625-687	4.94	.34	.02 *	3.52	.31	.01 *	1.42	.59	.01 *	51.98
<u>MOS</u>											
11B	116-121	4.83	.34		3.58	.32		1.25	.62		53.05
13B	128-147	4.82	.35		3.62	.33		1.21	.63		46.47
19E/K	34-41	4.97	.29		3.45	.24		1.53	.47		49.24
31C	91-94	4.93	.35		3.51	.30		1.43	.62		55.57
63B	98-107	4.86	.34		3.55	.29		1.31	.56		44.13
71L	98-109	4.97	.35		3.55	.37		1.42	.65		52.13
88M	132-141	4.83	.31		3.59	.31		1.25	.57		42.25
91A/B	89-100	4.92	.31		3.52	.28		1.40	.54		57.27
95B	117-136	5.11	.30	.08 *	3.41	.28	.04 *	1.69	.53	.06 *	62.61

* Significant at the .01 level.

^aThe rank order correlation between mean M-Eff. and mean AFQT across the nine MOS is .57.^bThe rank order correlation between mean L-Eff. and mean AFQT across the nine MOS is .66.^cThe rank order correlation between mean M-L Eff. and mean AFQT across the nine MOS is .60.^d11B, 13B, and 19E/K.^e31C, 63B, 71L, 88M, 91A/B, and 95B.

Preliminary Results Concerning Construct Validity

The Dimensionality of SJT Scores. The item-level scores for each of the three most promising scoring procedures were intercorrelated and factor analyzed using principal factor analysis. From two to five factors were extracted for each scoring procedure and rotated to a varimax solution. The results of these factor analyses did not generate any clearly definable dimensions, and were for the most part uninterpretable. The partially identifiable factors that did emerge in a few of these analyses involved (a) disciplining when appropriate, (b) avoiding disciplining when inappropriate, and (c) assigning work tasks effectively.

Respondents' scores for each response alternative (i.e., 0 if they chose that alternative, 1 if they did not) were intercorrelated and factor analyzed using the same procedures. From three to ten factors were extracted for each of the three scoring procedures. The factors obtained were defined primarily by the effectiveness of the response alternatives; effective alternatives had positive loadings and relatively ineffective alternatives had negative loadings. The factor loadings were otherwise uninterpretable.

It is important to note that this factor analysis of the response alternative level scores violates one of the assumptions on which factor analysis is based, independence of the scores that are factor analyzed. For example, if respondents chose response alternative "a" for the first item, by definition they could not also have chosen response alternative "b" for that same item. In general, however, the factor analysis results do not indicate that the SJT items and response alternatives can be divided into meaningful subconstructs or dimensions.

Estimates of Supervisory Behaviors Assessed by the SJT. Table 5.42 shows that the expert judgments of which supervisory behaviors are involved in the SJT response alternatives were very reliable. For each of the 10 dimensions, the interrater reliability of the mean rating (across all nine raters) ranges from .82 to .99, with a median of about .94. Reliabilities were particularly high for Referring and for Interacting Assertively With Superiors. It was apparently more obvious to the raters when response alternatives involved these behaviors. Reliabilities were lowest for Reasoning With Soldiers and for Communicating With Subordinates.

The extent to which the SJT measures performance related to each of these dimensions was assessed by computing the mean, across all response alternatives, of the mean number of points the nine raters assigned to each dimension. The second column of Table 5.42 presents these overall values. These means are highest for Gathering Information/Monitoring and Disciplining and lowest for Referring, Reasoning with Soldiers, and Giving Orders, but all 10 of the dimensions appear to be covered to a reasonable degree.

Table 5.42**Situational Judgment Test: Interrater Reliabilities for Response Alternative Dimensional Ratings, and Mean Rating and Mean Effectiveness for Each Dimension**

Dimension	Interrater Reliability ^a	Mean Rating Across All Response Alternatives	Mean Effectiveness
Referring	.99	.54	4.28
Interacting assertively with superiors	.97	1.13	4.44
Counseling	.93	1.04	4.49
Encouraging	.92	.90	4.10
Disciplining	.96	1.43	3.55
Gathering information/monitoring	.96	1.46	4.70
Reasoning with soldiers	.82	.61	3.87
Giving orders	.92	.66	3.89
Assigning tasks	.94	1.18	3.89
Communicating with subordinates	.86	.85	4.06

^aInterrater reliabilities are for the mean across nine raters.

Approximately how these dimensions correspond to the dimensions identified in the job analysis (Campbell, 1989) is shown in Table 5.43. Two of the dimensions identified in the job analysis, Act as a Model and Training Subordinates, do not correspond to any SJT dimensions. For Act as a Model, this is probably because the SJT is a maximal performance test, and by its very nature this dimension will probably be better tapped by measures of typical performance (e.g., performance ratings).

Table 5.43**Correspondence of Situational Judgment Test Dimensions With Job Analysis Dimensions**

Dimensions Identified in the Second-Tour Job Analysis	Dimensions Identified in the SJT Categorization
Planning Operations	Assigning Tasks
Directing/Leading Teams	Giving Orders
Monitoring/Inspecting	Gathering Information/Monitoring
Individual Leadership	Encouraging
	Reasoning With Soldiers
Acting as a Model	
Counseling	Counseling
	Disciplining
Communicating With Subordinates, Peers, and Supervisors	Communicating With Subordinates
	Interacting Assertively With Superiors
Training Subordinates	
Personnel Administration	Referring

The mean effectiveness of each of the 10 SJT dimensions across all of the SJT response alternatives was computed using the formula presented in Table 5.44. For each response alternative, the mean effectiveness rating from the experts was weighted by the extent to which that response alternative was judged to tap a particular dimension. These weighted effectiveness ratings were then added together for each dimension. Finally, to place all of the dimensions on the same metric, these dimension scores were divided by the sum (across all response alternatives) of the extent to which the SJT response alternatives tap the relevant dimension.

This procedure resulted in effectiveness scores for each of the 10 dimensions that are on the same 1 to 7 scale as the effectiveness ratings from the experts, and these scores are shown in the third column on Table 5.42. This analysis shows that response alternatives involving Gathering Information/Monitoring and Counseling tend to be more effective, and those involving Disciplining and Reasoning With Soldiers tend to be less effective.

Table 5.44

Situational Judgment Test: Computation of the Mean Effectiveness of Each Dimension

		143	
		\sum (resp. alt. effectiveness ^a) (resp. alt. dimension rating ^b)	
		$i = 1$	
Mean Effectiveness	=		
Indexes for Each Dimension		143	
		\sum	(resp. alt. dimension rating ^b)
		$i = 1$	

^aThe effectiveness rating from USASMA experts.

^bFor each of the 10 dimensions, the mean dimension rating across all nine raters.

During development of the SJT, an effort was made to develop response alternatives involving all of the different types of supervisory behaviors identified in the job analysis, and also to develop items for which a variety of different types of supervisory behavior would be the most effective. The results presented in Table 5.42 indicate that this effort was reasonably successful. However, because SJT development focused on capturing all of these dimensions, as opposed to reflecting the importance of the various dimensions for the job, these results should not be interpreted as reflecting the importance or effectiveness of each of the 10 dimensions for second-tour MOS.

Relationship of SJT Scores to Experience and Training. The mean SJT scores of soldiers who reported various levels of supervisory training are shown in Table 5.45. Soldiers who have attended no supervisory school scored almost half a standard deviation lower than those who have attended one or more supervisory schools. The first school attended is the Primary Leadership Development Course (PLDC), and the next level of supervisory training is the Basic Noncommissioned Officer Course (BNCOC). Mean scores were computed for soldiers who had attended PLDC only, and for those who had attended both PLDC and BNCOC. The latter group scored higher than the former, but the difference was small.

One potential confound in all of these comparisons is that the opportunity to attend these schools may be influenced by the individual's effectiveness as soldier or as supervisor. As a result, it is possible that these mean SJT score differences occurred because the more effective soldiers were given the opportunity to attend supervisory training. However, regardless of whether these mean score differences result from differential opportunities or training in the relevant supervisory skills, they support the construct validity of the SJT as a measure of supervisory skill.

Soldiers in the CVII sample were also asked to report how often they are required to supervise other soldiers, and mean SJT scores for soldiers subgrouped by their response to this question are reported in Table 5.45. For all three SJT scoring procedures, the expected pattern was found; higher levels of supervisory responsibility are associated with higher SJT scores.

Table 5.45

Situational Judgment Test: Mean Scores for Soldiers With Different Levels of Supervisory Training and Experience

	N	M-Eff.	L-Eff.	M-L Eff.
Attended one or more supervisory schools	560-603	4.97	3.50	1.47
Attended no supervisory school	327-371	4.81	3.62	1.20
Attended PLDC	477-515	4.96	3.51	1.46
Attended PLDC and BNCOC	81-84	4.99	3.46	1.53
<i>How often required to supervise other soldiers:</i>				
Never	87-99	4.87	3.63	1.23
Sometimes fill in for regular supervisor	294-327	4.86	3.58	1.29
Often fill in for regular supervisor	125-135	4.90	3.53	1.38
Regularly supervise other soldiers	391-415	4.96	3.49	1.47

The largest differences are for the L-Effectiveness score; soldiers who report that they regularly supervise other soldiers obtain L-Effectiveness scores almost half a standard deviation better (i.e., lower) than those of soldiers who report that they never supervise other soldiers. The smallest difference is for the M-Effectiveness score.

These results for supervisory experience are slightly different than those for supervisory training, where the largest mean differences were found for the M-Effectiveness score. Perhaps this is because supervisory experience sometimes involves making mistakes and learning the consequences of these mistakes (i.e., learning to identify ineffective responses), but supervisory training is more likely to focus on identifying effective supervisory responses.

Correlations between the extent of self-reported supervisory responsibilities (on a scale of 1 to 4) and each of the three SJT scores were also computed and are reported on Table 5.46. These correlations are fairly low, but all are significantly different from zero. Table 5.46 also shows the correlations between self-reported time in a supervisory position and these same three SJT scores. These correlations are also fairly low, but all are significant. Thus, the results for all three scoring procedures indicate that amount of experience has a small positive relationship with SJT scores.

Table 5.46**Correlations Between Situational Judgment Test Scores and Supervisory Experience^a**

	N	M-Eff.	L-Eff.	M-L Eff.
Time in supervisory position	757-820	.13	-.14	.14
How often required to supervise (1 to 4 scale)	897-981	.13	-.15	.15

^aA correlation of .07 is significant at the .05 level.

Conclusions and Recommendations

The results of the SJT data analyses indicate that the measure has appropriate distributional characteristics for the CVII sample. The five scoring procedures all resulted in scores with reasonable variance and internal consistency reliabilities and item-total correlations were quite high. Based on the psychometric characteristics, the most promising score appears to be M-L Effectiveness, which has an internal consistency reliability of .75.

Preliminary investigations on the construct validity of the SJT provided some evidence that the SJT is a valid measure of supervisory job knowledge. The SJT response alternatives describe supervisory behaviors that cover most of the range of behaviors identified in an earlier job analysis. The mean scores of soldiers with various levels of supervisory training and experience seem to indicate that the knowledges measured by the SJT are, to some extent, learned on the job. Finally, the factor analysis results indicate that the SJT can be seen as measuring a fairly unidimensional construct. Perhaps this construct could indeed be called "supervisory judgment."

The correlations of SJT scores with other second-tour criterion measures (reported in Chapter 6) also provide some support for the construct validity of the SJT as a measure of supervisory job knowledge. SJT scores correlate moderately with scores on the supervisory simulations and with the rating composite called Leading/Supervising. Correlations of SJT scores with scores on the job knowledge tests are higher, but this is not surprising in view of the fact that these all are paper-and-pencil tests.

On the basis of the present data analysis results, the M-L Effectiveness score appears to be a good summary of the information contained in the SJT. This score does, however, combine two somewhat different types of information: the ability to identify the most effective response to a situation and the ability to identify the least effective response. The M-Effectiveness and L-Effectiveness scores correlate quite highly, but the relationships with levels of supervisory training and experience are slightly different for these two scores. Thus, further exploration of the construct validity of the M-Effectiveness and L-Effectiveness scores separately is warranted.

If the SJT is used in future data collections, replacement of the three items with extremely low item-total correlations should be considered. Many items available from the SJT development research could be used for this purpose, and it is recommended that three of these items be included in future versions of the SJT.

NCO Supervisory Simulation Exercises

This section outlines procedures used to develop basic criterion scores for the three leadership/supervisor role-play exercises developed during Project A to measure components of second-tour (NCO) performance. These measures were developed to assess NCO performance in important job areas that were judged to be best assessed through the use of interactive exercises. The simulation exercises were designed to assess performance in the areas of counseling subordinates and training subordinates. A trained evaluator (role player) played the role of a subordinate to be counseled or trained and the examinee assumed the role of a first-line supervisor who was to conduct the counseling or training. Evaluators also scored the examinees' performances using a standard set of rating scales.

Simulation Exercise Content

Here are brief descriptions of the three simulation exercises:

- **Personal Counseling Simulation** --

Supervisory Problem: A PFC is exhibiting declining job performance and personal appearance. Recently, the PFC's wall locker was left unsecured. The supervisor has decided to counsel the PFC about these matters.

Subordinate Role: The soldier is having difficulty adjusting to life in Korea and is experiencing financial problems. The role player is trained to initially react defensively to the counseling but to calm down if the supervisor handles the situation in a non-threatening manner. The subordinate will not discuss personal problems unless prodded.

- **Disciplinary Counseling Simulation** --

Supervisory Problem: There is convincing evidence that the PFC lied to get out of coming to work today. The PFC has arrived late to work on several occasions and has been counseled for lying in the past. The PFC has been instructed to report to the supervisor's office immediately.

Subordinate Role: The soldier's work is generally up to standards which leads the soldier to believe that he or she is justified in occasionally "slacking off." The subordinate had slept in to nurse a hangover and then lied to cover it up. The role player is trained to initially react to the counseling in a very polite manner but to deny that he or she is lying. If the supervisor conducts the counseling

effectively, the subordinate eventually admits guilt and begs for leniency.

● **Training Simulation** --

Supervisory Problem: The commander will be observing the unit practice formation in 30 minutes. The PVT, although highly motivated, is experiencing problems with the hand salute and about face.

Subordinate Role: The role player is trained to demonstrate feelings of embarrassment that contribute to the soldier's clumsiness. Training also includes making very specific mistakes when conducting the hand salute and about face.

For each exercise, examinee performance was evaluated on 3-point rating scales reflecting specific behaviors tapped by the exercises and a 5-point overall effectiveness rating scale. A 5-point Overall Fairness rating and a 5-point Overall Affect rating were also provided for personal counseling and disciplinary counseling, respectively. Examples of two rating scale items from the personal counseling simulation are presented in Figure 5.4.

-
- | | |
|-------|--|
| _____ | 1. Develops rapport at the start of the session. |
| | 3 = Opens the interview in a pleasant, non-threatening manner. |
| | 2 = Opens the interview in a generally non-threatening manner but uses a tone of voice or non-verbal actions that leave the subordinate feeling somewhat defensive. |
| | 1 = Opens the interview in a hostile or threatening manner, leaving the subordinate feeling very defensive from the start. |
| _____ | 2. States the purpose of the counseling session clearly and concisely. |
| | 3 = Outlines all topics to be covered (e.g., the purpose is to discuss the wall locker that was left open last night, any problems the subordinate may be having, what might be done to resolve them, etc.). |
| | 2 = States at least one general topic to be discussed (e.g., says the purpose is to talk about the subordinate's recent poor performance). |
| | 1 = Fails to state a purpose for the session; instead jumps directly into problems. |
-

Figure 5.4. Example rating items from Personal Counseling Simulation.

CVII Simulation Sample Data

The basic scores for the supervisory simulations were developed using data from the CVII sample. The features of the sample and the data collection that are particularly important for the role-play exercises are described below.

A total of 18 individuals were selected to serve as role players/scorers. To represent first-tour soldiers as closely as possible, the role players were generally under 25 years old and had prior experience in military service. The majority (14 of the 18) were male. More men were selected intentionally because there were no women in the three combat jobs (11B, 13B, and 19E/K). Accordingly, it was felt that making soldiers counsel or train women when they did not work closely with women might adversely affect their performance in the simulation exercise. Seven of the role players were black; the remainder were white.

Role Player/Scorer Training. All role players participated in a 3-day training session to learn how to play the subordinate roles and score the exercises. This training consisted of oral review of the exercises, practice in playing the roles and receiving feedback, and practice in rating performance in the role plays and receiving feedback. Although all individuals were trained on all three of the exercises, at the end of training each scorer was assigned one exercise for which he or she would be primarily responsible.

Examinee Sample. Data were collected on a total of 976 second-tour soldiers, 720 of whom came from CONUS locations and 256 from USAREUR locations. The numbers of soldiers in each MOS were as follows: 11B, Infantryman, 117; 13B, Cannon Crewman, 154; 19E, Armor Crewman, 31; 19K, Armor Crewman, 9; 31C, Single Channel Radio Operator, 93; 63B, Light Wheel Vehicle Mechanic, 110; 71L, Administrative Specialist, 99; 88M, Motor Transport Operator, 129; 91A, Medical Specialist, 85; and 95B, Military Police, 138.

Data Collection Procedures. One team of three scorers was assigned to each CONUS location and one team of five scorers was assigned to cover all of the USAREUR locations. Scorers were assigned to data collection locations based on their availability. As mentioned above, each scorer was assigned one primary exercise to administer and score, but in a few instances scorers had to administer and score an exercise other than their primary exercise. This occurred relatively infrequently and usually because the primary scorer for an exercise was unavailable for a given day of the data collection. On average, 4.87 different scorers evaluated soldiers in each of the nine MOS.

Simulation Exercises Score Analyses

Descriptive Statistics. Descriptive statistics (means and standard deviations) for the individual rating items and overall ratings are presented in Table 5.47. Also presented are one-rater reliabilities that were computed based on a sample of approximately 70 ratees in USAREUR. Shadow scoring could be conducted in USAREUR because enough raters were available. Since only three role player/scorers were assigned to each CONUS location, it was not possible to collect multiple ratings of CONUS examinees.

Table 5.47**Descriptive Statistics for Simulation Exercises**

	Personal Counseling	Disciplinary Counseling	Training
Item Ratings			
<u>N</u> Rating Items	20	13	12
<u>N</u> ratees	974	976	956
Median Mean Rating ^a	2.06	2.22	2.41
Range	1.41-2.89	1.71-2.67	2.16-2.85
Median Standard Deviation	.69	.73	.69
Range	.36-.80	.49-.82	.43-.81
Item Reliabilities			
<u>N</u> Ratees Scored by two raters	64	74	70
Median One-Rater Reliability	.68	.78	.68
Range	.00-.89	.20-.87	.58-.95
Overall Effectiveness			
Mean ^b	2.65	2.66	2.86
Standard Deviation	1.26	1.00	1.14
One-Rater Reliability	.84	.76	.89
Overall Affect			
Mean ^b	3.05		
Standard Deviation	1.12		
One-Rater Reliability	.77		
Overall Fairness			
Mean ^b		3.09	
Standard Deviation		.94	
One-Rater Reliability		.81	

^aOn a 3-point scale.^bOn a 5-point scale.

Overall, the means and standard deviations of the ratings were within the expected range. Means of individual items on a 3-point scale ranged from 1.41 to 2.89, with median mean ratings ranging between 2.06 and 2.41 for the three exercises. The standard deviations for individual items ranged from .36 to .82, with median standard deviations of approximately .70. For the overall scales (i.e., overall effectiveness, overall affect, and overall fairness), the mean ratings ranged from 2.65 to 3.09 on 5-point scales, with the standard deviations of these ratings ranging from .94 to 1.26.

The one-rater reliabilities were uniformly high, with median reliabilities ranging from .68 to .78 for the three exercises. For the overall ratings, reliabilities were also uniformly high, ranging from .76 to .89.

Factor Analyses. Principal factor analyses with orthogonal and oblique rotations were performed for each exercise and the orthogonal versus oblique solutions were virtually identical. Separate factor analyses were performed for the CONUS versus USAREUR data because one team of five scorers administered the exercises in all of the USAREUR locations whereas several different teams of scorers administered the exercises in CONUS. However, the factor analysis results were virtually identical between CONUS and USAREUR, so results will be reported based on the entire sample.

Factor analyses of the item ratings for all three exercises yielded three distinct factors reflecting each of the original exercises. Raters were assigned within exercises and the content dimensions could not overcome this source of method variance across factors.

Presented in Table 5.48 are the factor analysis results for the personal counseling exercise. Two factors resulted and were named "content" and "process". The content scale items reflected specific actions by the examinee during the counseling session, such as providing advice to the subordinate, offering assistance to help solve problems, and encouraging the subordinate to perform effectively in the future. The process scale contained items that more generally reflected the overall manner in which the examinee conducted the session, such as conducting the counseling in a professional, non-threatening manner, maintaining open communication, and developing rapport. Two items were omitted from the scales because they were judged to reflect process but loaded higher on the content scale.

The factor analysis results for the disciplinary counseling exercise are shown in Table 5.49. Two factors also resulted for this exercise and were named "content" and "interpersonal skills." Like the personal counseling exercise, the content scale for disciplinary counseling items reflected specific actions by the examinee during the disciplinary counseling session. Examples include stating the exact provisions of the punishment, and determining an appropriate corrective action. The interpersonal skills scale contained items that reflected the overall manner in which the examinee conducted the session, such as conducting the counseling in a professional manner, and diffusing rather than escalating potential arguments. Four items were omitted from the final two scales because they had similar loadings on both factors but did not form their own factor when more factors were extracted.

Table 5.48

Personal Counseling Exercise Items and Factor Analysis Results^a

	Factor 1	Factor 2	h^2 ^b
<u>Content Scale</u>			
12. Provides advice to the subordinate concerning actions that should be taken to solve problems.	<u>.71</u>	.00	.50
19. At the end of the session, summarizes what has been discussed and decided.	<u>.65</u>	-.19	.46
13. Offers assistance to help solve problems.	<u>.64</u>	.02	.41
18. Encourages the subordinate to perform effectively in the future.	<u>.64</u>	-.09	.42
20. Sets a time or date to follow up with the subordinate.	<u>.61</u>	-.20	.41
9. Asks open-ended, fact-finding questions that uncover important and relevant information.	<u>.55</u>	.24	.36
3. Gives the subordinate positive feedback for his/her overall good past performance.	.51	.07	.27
2. States the purpose of the counseling session clearly and concisely.	<u>.50</u>	.00	.25
6. Avoids jumping to premature conclusions.	<u>.48</u>	.24	.29
4. Explains what was expected of the subordinate and how s/he failed to meet standards.	<u>.47</u>	-.15	.24
11. Focuses on one problem at a time; avoids jumping from topic to topic.	<u>.37</u>	.14	.16
<u>Process Scale</u>			
16. Maintains a professional body posture.	-.15	<u>.57</u>	.35
15. Maintains eye contact during the interview.	-.02	<u>.54</u>	.29
5. Conducts the counseling in a professional manner.	.16	<u>.53</u>	.31
14. Maintains open communication.	.01	<u>.52</u>	.27

(continued)

Table 5.48 (Continued)

Personal Counseling Exercise Items and Factor Analysis Results^a

	Factor 1	Factor 2	h^2 ^b
<u>Process Scale</u> (continued)			
10. Does not allow the subordinate to take control of the counseling session.	-.04	<u>.51</u>	.26
1. Develops rapport at the start of the session.	.03	<u>.46</u>	.21
8. Does not interrupt the subordinate when s/he is talking.	-.18	<u>.45</u>	.23
<u>Items Omitted</u>			
17. Maintains a supportive attitude and displays interest throughout the session.	<u>.55</u>	.22	.35
7. Encourages the subordinate to actively participate in the counseling session.	<u>.41</u>	.36	.30
Eigenvalue	4.07	2.27	6.34

^aPrincipal factor analysis with oblique rotation.

^b h^2 = communality (sum of squared factor loadings) for variables.

Table 5.49

Disciplinary Counseling Exercise Items and Factor Analysis Results^a

	Factor 1	Factor 2	h^2 ^b
<u>Content Scale</u>			
11. States the exact provisions of the punishment.	-.04	<u>.68</u>	.46
10. Determines an appropriate corrective action.	-.08	<u>.61</u>	.38
12. Makes sure the subordinate understands what has been discussed and decided.	.11	<u>.57</u>	.34
8. Remains focused on the immediate problems (i.e., the subordinate's absence and/or lying).	-.08	<u>.39</u>	.16
<u>Interpersonal Skills Scale</u>			
5. Conducts the counseling in a professional manner.	<u>.78</u>	-.13	.63
6. Diffuses rather than escalates potential arguments.	<u>.67</u>	-.17	.48
13. Counsels with a productive attitude.	<u>.58</u>	.13	.35
4. Allows the subordinate to present his/her view of the situation.	<u>.55</u>	-.03	.30
7. Asks open-ended, fact-finding questions that uncover important and relevant information.	<u>.45</u>	.25	.27
<u>Omitted Items</u>			
1. States the purpose of the counseling session clearly and concisely.	.29	.15	.11
2. Explains what was expected of the subordinate.	.38	.34	.26
3. Tells the subordinate the impact of his/her behavior.	.35	.38	.27
9. Maintains eye contact during the session.	.04	.23	.05
Eigenvalue	2.28	1.77	4.06

^aPrincipal factor analysis with oblique rotation.^b h^2 = communality (sum of squared factor loadings) for variables.

For the training exercise, the factor analysis indicated that the data could best be summarized by one factor. The items for the training exercise are presented in Table 5.50.

Table 5.50

Training Exercise Items

1. Presents an overview of what will be learned.
 2. Organizes and presents the training steps in a logical sequence.
 3. Provides verbal instruction to the trainee.
 4. Demonstrates the task steps for the trainee.
 5. Identifies and corrects the trainee's errors.
 6. Allows the trainee an opportunity to practice each movement required to perform the task.
 7. Provides specific on-the-spot correction to the trainee, as appropriate.
 8. Provides positive feedback to the trainee following good performance.
 9. Encourages the trainee when mistakes are made.
 10. Speaks in a clear, distinct, and understandable manner.
 11. Avoids distracting behaviors.
 12. Conducts the training in a professional manner.
-

To further investigate the appropriateness of a two-factor solution for the personal and disciplinary counseling exercises and a one-factor solution for the training exercise, a confirmatory factor analysis was conducted. The primary issue to be investigated for all three exercises was whether the data could best be summarized by one or two factors. Accordingly, the fit of one- and two-factor models was examined for each of the three exercises, and the resulting fit indexes, which include chi-square, adjusted goodness-of-fit (AGFI), and root mean square residual (RMS), are presented in Table 5.51.

As can be seen in the table, a two-factor solution for personal and disciplinary counseling seemed most appropriate. This is evidenced by decreases in the chi-square statistics and RMS indexes, and increases in the AGFIs when two-factor versus one-factor models were tested. For the training exercise, a one-factor model seemed to fit the data slightly better than a two-factor model and is certainly more parsimonious.

Table 5.51**Confirmatory Factor Analysis Results for Simulation Exercises**

Exercise	N Factors	df	χ^2	AGFI	RMS	χ^2/df
Personal Counseling	1	135	958.6	.860	.115	7.10
	2	134	829.1	.878	.102	6.19
Difference		1	129.5			
Disciplinary Counseling	1	27	498.6	.807	.154	18.47
	2	26	352.3	.858	.115	13.55
Difference		1	146.2			
Training	1	35	431.4	.856	.124	12.32
	2	34	428.3	.853	.126	12.60
Difference		1	3.1			

Finally, intercorrelations between the five scale scores (two for personal counseling, two for disciplinary counseling, and one for training) were computed, and alpha coefficients were also calculated for each of the scale scores. These values are presented in Table 5.52. Overall, the alpha coefficients are uniformly high, ranging from .65 to .85. The majority of the intercorrelations, which were corrected for attenuation, are in the .30s or above.

Table 5.52**Intercorrelations Between Scale Scores for Simulation Exercises**

	PC	PP	DC	DI	T
Personal Counseling - Content	(.83)				
Personal Counseling - Process	.53	(.69)			
Disciplinary Counseling - Content	.39	.06	(.65)		
Disciplinary Counseling - Interpersonal Skills	.30	.18	.46	(.75)	
Training	.33	.37	.30	.12	(.85)

Summary and Conclusions

This section has presented the procedures used to develop basic criterion scores for the simulation exercises. Scores were identified through the use of principal factor analyses within each exercise. To summarize the results of these analyses, two-factor solutions were chosen as the most psychologically meaningful and empirically defensible for the personal counseling and disciplinary counseling exercises.

The factors for the personal counseling exercise were named "content scale" and "process scale." The factors for the disciplinary counseling exercise were named "content scale" and "interpersonal skills scale." For the training exercise, a one-factor solution seemed most appropriate.

Army Job Satisfaction Questionnaire (AJSQ)

Prior to the 1988 data collection, the idea of administering a job satisfaction measure to all Project A soldiers was proposed. The rationale was that a job satisfaction measure would provide a more complete picture of person-job fit than that provided by job performance measures alone. It was anticipated that this information would also be particularly useful for predicting attrition.

The Army Job Satisfaction Questionnaire (AJSQ) measured six aspects of satisfaction: (a) supervision, (b) co-workers, (c) promotions, (d) pay, (e) work, and (f) Army. The form contained 20 items designed to tap these six components. The work and Army scales contain four items each, and the other scales have three items. Soldiers responded to the items using a 5-point scale which ranged from very dissatisfied to very satisfied.

The analyses reported herein are based on data from both first- and second-tour soldiers. Table 5.53 provides a description of this sample of 12,260 soldiers.

The results of a principal components analysis of the 20 AJSQ items are shown in Table 5.54. The rotated factor pattern (using varimax rotation) depicts a very clean solution confirming the distinctiveness of the six satisfaction components. The AJSQ subscores, then, are computed by summing the three to four items comprising each component. A composite score was computed by summing all 20 AJSQ items, and one item (Item #20) was identified as being a reasonable measure of overall satisfaction. This item is also part of the Army satisfaction subscore. It reads "In general, how satisfied are you with all aspects of Army life (including work, services, etc.)?"

The intercorrelations among the AJSQ subscores are shown in Table 5.55. Note that the high correlation between the Army satisfaction score and the overall satisfaction score is attributable to the fact that the overall score is based on one of the Army subscale items. Coefficient alphas for each of the AJSQ subscores are also shown in Table 5.55. The values range from .80 to .91, which indicates that these scores exhibit a reasonably high degree of internal consistency.

Table 5.53

Army Job Satisfaction Questionnaire (AJSQ) Sample Description

Sample Size: N = 12,260

Race: Black 27.5% (n = 3372) White 64% (n = 7844)
 Hispanic 4% (n = 469) Other 4.5% (n = 557)

Gender: Male 88% (n = 10801) Female 12% (n = 1427)

Paygrade:	<u>Percent</u>	<u>N</u>
E-1	1.5	180
E-2	9.5	1149
E-3	47	5699
E-4	38	4602
E-5	3	410
E-6	1	11

Primary MOS:	<u>Percent</u>	<u>N</u>	<u>Percent</u>	<u>N</u>	
11B	8	984	55B	2	271
12B	7	826	63B	7	833
13B	9	147	67N	2	196
16S	4	452	71L	6	773
19E	3	307	76Y	7	776
19K	6	775	88M	8	917
27E	1	90	91A	8	917
29E	1	110	94B	7	791
31C	5	613	95B	5	581
51B	2	207	96B	1	129
54B	4	491			

Table 5.54

AJSQ Principal Components Analysis^a

Item	Work	Army	Pay	Supervision	Promotion	Co-workers	h^2 ^b
01	.15	.10	.07	.89	.10	.05	.84
02	.13	.12	.06	.90	.10	.06	.86
03	.21	.16	.06	.85	.12	.09	.82
04	.10	.08	.05	.08	.08	.77	.62
05	.13	.07	.03	.05	.06	.84	.73
06	.14	.13	.07	.04	.09	.87	.81
07	.11	.07	.11	.10	.89	.08	.84
08	.20	.23	.13	.13	.82	.11	.81
09	.16	.17	.12	.11	.90	.09	.90
10	.11	.22	.89	.07	.12	.06	.88
11	.10	.19	.86	.07	.11	.06	.81
12	.11	.24	.88	.06	.12	.05	.86
13	.85	.15	.10	.14	.12	.11	.80
14	.87	.11	.08	.13	.12	.11	.82
15	.85	.09	.08	.12	.12	.12	.78
16	.74	.25	.10	.17	.14	.13	.69
17	.19	.68	.19	.11	.18	.13	.60
18	.11	.80	.17	.08	.08	.06	.70
19	.07	.80	.17	.12	.08	.07	.70
20	.24	.78	.19	.13	.18	.12	.77
Eigenvalue	3.10	2.76	2.53	2.52	2.51	2.20	15.64

^aVarimax rotation.^b h^2 = communality (sum of squared factor loadings) for variables.

Table 5.55**AJSQ Subscore Intercorrelation Matrix**

	Supervision	Co-workers	Promotion	Pay	Work	Army	Composite
Supervision	.90						
Co-workers	.20	.80					
Promotion	.30	.25	.91				
Pay	.21	.17	.32	.91			
Work	.38	.31	.38	.29	.90		
Army	.33	.28	.39	.48	.42	.84	
Composite	.62	.52	.67	.62	.74	.75	.90
Overall (Item 20)	.32	.27	.38	.42	.43	.87	.70

Note. Ns range from 11,886 to 12,164. Alpha coefficients are shown in the diagonal.

Descriptive statistics for each of the AJSQ subscores are shown in Table 5.56. The means and standard deviations are shown for the whole sample as well as separately by race and gender. Although there are differences in satisfaction levels across race and gender subgroups, none of the differences appear to be particularly large.

Summary of Second-Tour Criterion Score Development

At this point in the Career Force Project, the second-tour performance data, which were collected on a total CVII sample of 1053 soldiers at the E-4 and E-5 level, have been edited, scored, and subsequently aggregated into a set of scores referred to as the "basic criterion scores." There are 22 such scores, not counting the Army Job Satisfaction Questionnaire. The 22 scores represent the project's best effort to capture the valid information in each measure in the most informative way possible. A summary list of these best scores provided by each measure is presented in Figure 5.5.

Table 5.56

AJSQ Means* and Standard Deviations by Race and Gender

	Supervision		Co-Workers		Promotion		Pay		Work		Army		Composite	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
All	9.49	3.27	9.89	2.62	7.73	3.36	7.83	3.05	11.25	4.24	10.12	3.61	56.29	13.36
Males	9.44	3.27	9.91	2.60	7.75	3.37	7.75	3.03	11.16	4.22	10.07	3.61	56.06	13.37
Females	9.78	3.31	9.69	2.74	7.58	3.26	8.52	3.08	11.81	4.30	10.46	3.59	57.84	13.19
White	9.51	3.26	9.74	2.64	7.56	3.35	7.83	3.02	10.95	4.28	9.97	3.59	55.54	13.19
Black	9.45	3.34	10.24	2.56	8.02	3.40	7.75	3.12	11.82	4.11	10.32	3.65	57.60	13.67
Hispanic	9.62	3.13	9.86	2.53	8.32	3.28	8.46	3.00	12.10	3.98	10.94	3.69	59.30	13.48

*Responses are on a 1 - 5 scale. The supervision, co-workers, promotion, and pay subscales consist of three items each. The work and Army subscales consist of four items each.

Hands-On Job Sample Test

MOS-specific task performance score
Common task performance score

Job Knowledge Test

MOS-specific task knowledge score
Common task knowledge score

Army-Wide Rating Scales

Leadership/supervision composite
Technical proficiency and effort composite
Maintaining personal discipline composite
Physical fitness and military bearing composite

MOS-Specific BARS Scales

One overall MOS BARS composite score

Combat Performance Prediction Scales

One overall combat scale composite

Personnel File Form

Awards and Commendations
Articles 15/Flag Actions
Physical Qualification
M16 Qualification
Number of Military Training Courses
Promotion Rate

Situational Judgment Test

One score obtained by subtracting the total "ineffectiveness" score from the total "effectiveness" score

Supervisory Simulation Exercises

Personal Performance Counseling: Interaction Process
Personal Performance Counseling: Interaction Content
Discipline Problem Counseling: Interaction Process
Discipline Problem Counseling: Interaction Content
One-on-One Training: Total Composite Score

Figure 5.5. Summary list of second-tour basic criterion scores.

The next step in the analysis of second-tour performance is to use the 22 basic scores as input for an effort to "model" the basic structure of second-tour (NCO) performance.² At this point, each of the scores is specific to a particular measurement method and there are still too many scores to be used as criterion measures for validation purposes. Consequently, a critical issue concerns how the total covariation in the 22 basic scores can be best represented by a smaller number of basic performance factors. To phrase it another way, what is the latent structure of performance for the second-tour NCO?

The next chapter describes the project effort to use the 22 basic criterion scores and the CVII data to model the basic nature of second-tour performance.

²The Combat Performance Prediction Scale was not administered to women during CVII data collection; therefore, it was not used in the model building exercise described in Chapter 6.

Chapter 6

Modeling of Second-Tour Performance

The analysis of the basic criterion scores for second-tour performance was guided by the same conceptual framework as the development of performance factor scores for first-tour performance (Campbell, McHenry, & Wise, 1990). That is, total performance is assumed to be composed of a small number of distinct components such that aggregating them into one score covers up too much information about an individual's relative proficiency on the separate factors. The meaning of each separate component is independent (conceptually at least) of measurement method. The major components that are hypothesized to exist comprise the so-called latent structure of performance.

The analysis task is to determine which model (i.e., a particular specification for the number of components and their substantive content) of the latent structure best fits the observed data. A good fit implies that the composite scores used to measure each major component are both a parsimonious and a valid representation of the basic nature of performance.

The above approach portrays the basic structure in terms of a factor model that rules out a general factor, proposes no causal relationships among the components, and, at least initially, incorporates no hierarchical properties. However, the true correlations among the latent variables, the existence of "methods" factors, and the multidimensionality of the scores from individual criterion measures are matters for investigation and hypothesis testing.

THE INPUT DATA

The data to be analyzed were collected from the CVII sample using the measures of second-tour performance developed as part of Project A (Campbell, 1989). The specific individual criterion scores (herein called the "basic" scores) derived from applying these measures were described in the previous chapter. There are 22 such basic scores.¹ By the original project design these scores were intended to cover the entire performance domain (as specified by previous job analyses) with multiple measurement methods, if possible. A brief summary of how these scores came to be is as follows:

Hands-On Performance Tests. Analyses of the percent-go scores for the various hands-on task tests again suggested two overall clusters of tasks: MOS-Specific and General Soldiering. We examined the functional categories and the six-skill-category breakdown of hands-on tasks that were also explored in developing the first-tour performance model, but did not find sufficient agreement across MOS to warrant retention of more than two hands-on scores.

¹The list of 22 basic criterion scores given in Figure 5.5 was slightly modified for the development of the models described in this chapter. Because the Combat Performance Prediction Scale was not administered to women soldiers during the CVII data collection, it was not used in the performance models. An additional variable from the Personnel File Folder, the Skill Qualification Test (SQT) self-report score, was added to the scores used in analyzing the models but was not retained in the final listing.

Job Knowledge Tests. The job knowledge tests were also organized around a specific sample of tasks. Scores at the item and task level were analyzed to determine the dimensions that best summarized the information from these tests. Again, a two-factor model with separate general soldiering and MOS-specific task scores was judged best.

Army-Wide Performance Ratings. We focused on the supervisor BARS ratings, as they had greater face validity and were considerably more complete than the peer ratings. Four factors were identified: Leadership, Technical Proficiency, Discipline, and Physical Fitness. This result conformed nicely to original expectations that underlay the selection of the individual scales.

MOS-Specific Performance Ratings. We did not find any consistent structure within the MOS-specific BARS ratings and concluded that a single score provided an optimal summary of the information in these ratings.

Personnel File Folder Measures. Analyses of the items on the administrative records questionnaire and the supplemental data from the Enlisted Master File suggested seven overall scores: awards, disciplinary actions, training courses completed, grade deviation (promotion rate relative to the norm), physical readiness scores, marksmanship scores, and SQT scores. The training courses and SQT measures were additions to the scores identified in the first-tour modeling; the other five scores corresponded exactly to those identified in the first-tour.

Situational Judgment Test. A number of scoring alternatives for the Situational Judgment Test were examined. In the end, a single score using the effectiveness ratings for the options selected and combining the judgments of most and least effective (by taking the difference in rated effectiveness) was selected as the best measure.

Counseling and Training Simulation Exercises. Analyses of the rating items for the three role-play exercises used to assess counseling and training skill suggested that unique information was conveyed by separate scores for each of the three exercises. In addition, there was some clustering of items within the personal counseling and disciplinary counseling exercises into separate process and content categories. Five scores were thus identified for further analyses: Disciplinary Content, Disciplinary Process, Personal Counseling Content, Personal Counseling Process, and Training.

After scores from each measurement method were defined, the development of the overall second-tour performance model began with an examination of the correlations among the scores from the different measurement methods. These correlations are shown in Table 6.1. Exploratory factor analyses suggested five to six substantive factors, generally similar to those in the first-tour model, and also suggested methods factors for at least the ratings and the written measures.

The exploratory results were reviewed by project staff and several alternative models were suggested for "confirmation". Because the sample sizes were limited, it was not feasible to conduct split-sample cross-validation. The confirmation of initial results is thus primarily suggestive. Collection and analysis of second-tour data for the Longitudinal Validation sample will provide a much more definitive opportunity to confirm these initial results.

Table 6.1

Correlations Among the CVII Summary Measures Based on All Soldiers With Complete Data After Minimal Imputation

Criterion Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
	Awd	Art	TrC	Grd	SQT	PhR	M16	HOG	HOS	JKG	JKS	R-L	R-T	R-D	R-F	MOB	RDC	RDP	RCC	RCP	RT	SJT
Personnel File Folder																						
1 PFF-Awards	xx	-05	30	28	20	10	14	07	06	05	11	22	18	11	16	15	10	10	12	07	08	09
2 PFF-Art/Flg	-05	xx	-17	-17	-03	-13	-03	-04	-00	00	-05	-12	-09	-16	-17	-07	-03	-06	-08	-11	-09	-02
3 PFF-Courses	30	-17	xx	40	21	19	20	16	13	19	16	31	24	19	31	23	13	04	17	11	15	14
4 Grade Deviation	28	-17	40	xx	15	18	14	15	14	17	18	31	20	18	26	21	16	04	13	07	15	20
5 PFF-SQT	20	-03	21	15	xx	03	11	16	20	26	25	18	20	12	08	22	12	04	06	08	13	19
6 PFF-Phys Read	10	-13	19	18	03	xx	09	-01	01	-05	-06	13	12	09	33	08	05	-04	08	07	10	-06
7 PFF-M16/19	14	-03	20	14	11	09	xx	11	11	08	01	13	11	03	12	14	-02	-14	-03	-00	09	03
Hands-On Tests																						
8 H0-General	07	-04	16	15	16	-01	11	xx	17	30	20	08	07	06	05	07	10	04	12	08	15	09
9 H0-Job Specific	06	-00	13	14	20	01	11	17	xx	23	42	08	11	08	08	09	07	06	02	04	10	11
Job Knowledge Tests																						
10 JK-General	05	00	19	17	26	-05	08	30	23	xx	48	13	17	10	08	15	06	03	16	11	21	40
11 JK-Job Specific	11	-05	16	18	25	-06	01	20	42	48	xx	13	13	11	04	12	10	11	18	12	14	34
Supervisor Ratings																						
12 AWB-Leadership	22	-12	31	31	18	13	13	08	08	13	13	xx	82	70	61	74	06	03	09	15	16	17
13 AWB-Tech Prof	18	-09	24	20	20	12	11	07	11	17	13	82	xx	70	59	75	04	-00	03	14	13	15
14 AWB-Discipline	11	-16	19	18	12	09	03	06	08	10	11	70	70	xx	56	63	-01	-01	03	17	09	16
15 AWB-Phys Fitness	16	-17	31	26	08	33	12	05	08	08	04	61	59	56	xx	53	02	-02	08	16	14	09
16 MOB-MOS Total	15	-07	23	21	22	08	14	07	09	15	12	74	75	63	53	xx	03	-05	01	10	14	11
Simulation Exercises																						
17 RP-Disc Content	10	-03	13	16	12	05	-02	10	07	06	10	06	04	-01	02	03	xx	39	32	11	26	07
18 RP-Disc Process	10	-06	04	04	04	-04	-14	04	06	03	11	03	-00	-01	-02	-05	39	xx	29	20	16	09
19 RP-Coun Content	12	-08	17	13	06	08	-03	12	02	16	18	09	03	03	08	01	32	29	xx	43	32	18
20 RP-Coun Process	07	-11	11	07	08	07	-00	08	04	11	12	15	14	17	16	10	11	20	43	xx	30	14
21 RP-Training Tot	08	-09	15	15	13	10	09	15	10	21	14	16	13	09	14	14	26	16	32	30	xx	12
Situational Judgment Test																						
22 SJT-Most/Least	09	-02	14	20	19	-06	03	09	11	40	34	17	15	16	09	11	07	09	18	14	12	xx

Note. N = 1,009. Decimals have been omitted.

The initial results distinguished five factors that corresponded closely to those identified in the first-tour performance model. Alternative models developed by project staff concerned the possible identification of a sixth factor that separated aspects of leadership and supervisory judgment from the previous Effort and Leadership factor and also concerned the loadings of the new Simulation and SJT measures on this sixth factor or on the existing factors.

PROCEDURE

LISREL VI (Joreskog & Sorbom, 1986) was used to estimate the parameters and evaluate the fit of each of the alternative models. In this program, confirmatory factor analysis parameters are organized into three matrixes:

(1) The factor loadings, modeled with the Lambda Y matrix, give the regressions of each observed score on the underlying factors. This matrix is tightly constrained, with each observed variable loading on only one or two factors.

(2) The Psi matrix specifies the correlations among the underlying factors. Methods factors were constrained to be uncorrelated with each other and with each of the substantive factors. This means that all of the "cross-method" correlation had to be explained by common loadings on substantive factors and by intercorrelations among the substantive factors.

(3) The final matrix, Theta Epsilon, gives the variances and covariances among the unique components of each of the observed variables. The unique components of the observed variables represent the information that would be lost if the data were summarized by scores on the underlying factors and so were treated as measurement error. This means that the covariances across measures, represented by the off-diagonal elements of Theta Epsilon, are constrained to be zero and only the diagonal elements (the unique variances) are left to be estimated.

The LISREL VI program provides a number of fit statistics that can be used in assessing hypotheses about the data. First, there is an overall chi-square fit statistic that can be used to test the hypotheses that the overall correlation matrix differs from the best-fitting model-based matrix only by sampling error. Deviations from multivariate normality among the observed variables will affect the appropriateness of this statistic and, for relatively large sample sizes, the fit may be affected by factors that are not of practical significance.

A second type of statistic results from the comparison of the chi-square fit statistics for nested models. This allows for a test of the significance of improvement in fit as additional parameters are introduced. The program provides "modification indexes" which estimate the improvement in chi-square fit that would result if a particular constrained parameter were set free (and all other constraints remained in place). This allows for a test of the nonsignificance of parameters not included in the model.

Finally, the program also provides "t-value" statistics for each of the parameters in the model that can be used to test the difference of these parameters from zero. Again, this test assumes that all other parameters remain fixed.

The major models evaluated with the LISREL VI program were:

- (1) First-Tour Model: Includes five substantive and two methods factors, with the SJT and Simulation variables all loading on the Effort and Leadership factor.
- (2) Leadership Factor Model: Includes a sixth substantive factor with the SJT, Simulation, and Leadership Rating factor variables all loading on this factor. This model was evaluated with and without a separate Role-Play "methods" factor.
- (3) Training and Counseling Factor Model: Includes a sixth substantive factor with just the Simulation variables. No separate Role-Play methods factor could be estimated under this model.

Within each of these major models, a number of variations were explored. These variations included the loading of specific Personnel File Folder measures (e.g., the SQT, which was a new addition to the analyses, and the marksmanship scores, which did not fit well in the first tour) on Proficiency, Effort, Leadership, and Discipline factors. Loadings for the technical ratings on the proficiency factors also were explored as alternatives.

RESULTS

Of the three models, the Training and Counseling Factor Model came closest to fitting the observed data. The basic problem was that the Simulation exercise scores showed a good deal of internal consistency, but had very low correlations with any of the other performance measures. Consequently, any model that included a factor with loadings for both Simulation variables and other performance variables did not provide a reasonable fit to the data (as either the consistency among the Simulation exercises was underestimated or their correlations with other measures were overestimated).

Comparisons of Models

The parameter estimates for the fit to the Training and Counseling Model are shown in Table 6.2. After review of several minor variations, it was decided to drop the SQT and Marksmanship scores from the modeling and to exclude loadings for any of the ratings variables or the SJT on the proficiency factors. The overall fit statistics (see Table 6.3) were a chi-square of 374.9 with 149 degrees of freedom, an adjusted goodness-of-fit index of .948, and a root mean square residual of .055, indicating good, but not perfect, agreement with the empirical data.

LISREL Results for Training and Counseling Factor Model: Parameter Estimates

342

Table 6.3**LISREL Results for Training and Counseling Factor Model: Fit Statistics**

Measures of Goodness-of-Fit for the Whole Model:

Chi-Square with 149 Degrees of Freedom	374.9 (Prob. Level = 0.000)
Goodness-of-Fit Index	.963
Adjusted Goodness-of-Fit Index	.948
Root mean square residual	.055

Table 6.4 shows the *t*-value statistics for each of the estimated parameters, indicating that each was significantly different from zero. Table 6.5 shows the differences between the observed and fitted correlations (the residuals).

In order to explore the practical consequences of the misfit of the basic model, we freed additional parameters until a statistically acceptable chi-square was achieved. Tables 6.6 through 6.9 show the result of this process. (We refer to the resulting model as the "Overfit Model.")

In retrospect, some of the additional parameters were plausible and might have been included in the initial model. Specifically, we allowed for correlated errors among some of the Simulation exercise variables. This was reasonable because the two scores from a single exercise were based on a single set of raters, and because the correspondence between the "process" scores for the disciplinary and personal counseling exercises was not otherwise explained. Two of the Simulation variables were allowed to load on other factors (the training exercise score on General Soldiering and the personal counseling process score on Disciplinary) and had small, but significant, positive loadings.

Other parameter changes were less plausible. The awards and leadership ratings variables had slightly negative loadings on the General Soldiering factor, and the awards variable also had a slightly negative loading on the Disciplinary factor. Finally the physical readiness scores had a negative loading on the written methods factor while the personal counseling exercise scores had a small but positive loading on the ratings methods factor.

The general conclusion drawn from examination of the "overfit" model was that the additional parameters were all quite small and not of practical significance for defining overall criterion scores. The loadings were generally less than .2 in comparison to loadings of .4 to .8 for the originally hypothesized parameters. (The estimated within-exercise error correlations and the loading of awards on Personal Discipline were slightly greater than .2, but all less than .25.)

Table 6.4

LISREL Results for Training and Counseling Factor Model: t -Values

Lambda Y (t -values for factor loadings)									
	CT Job Spec	GP Gen Pref	EA Effort	PD Pers Disc	PF Phys Fit	TC Train	Written	Ratings	
PFF-Awards	.00	.00	12.00	.00	.00	.00	.00	.00	
PFF-Art/Flag	.00	.00	.00	8.12	.00	.00	.00	.00	
PFF-Courses	.00	.00	17.66	.00	.00	.00	.00	.00	
Grade Deviation	.00	.00	17.95	.00	.00	.00	.00	.00	
PFF-Phys Read	.00	.00	.00	.00	10.80	.00	.00	.00	
MO-General	.00	9.78	.00	.00	.00	.00	.00	.00	
MO-Job Specific	10.74	.00	.00	.00	.00	.00	.00	.00	
JK-General	.00	10.89	.00	.00	.00	.00	8.58	.00	
JK-Job Specific	13.10	.00	.00	.00	.00	.00	7.79	.00	
AJB-Leadership	.00	.00	13.27	.00	.00	.00	.00	30.92	
AJB-Tech Prof	.00	.00	9.30	.00	.00	.00	.00	32.01	
AJB-Discipline	.00	.00	.00	-8.99	.00	.00	.00	24.73	
AJB-Phys Fitness	.00	.00	.00	.00	14.19	.00	.00	17.63	
MOB-MOS Total	.00	.00	8.09	.00	.00	.00	.00	27.19	
RP-Disc Content	.00	.00	.00	.00	.00	15.30	.00	.00	
RP-Disc Process	.00	.00	.00	.00	.00	13.13	.00	.00	
RP-Coun Content	.00	.00	.00	.00	.00	19.37	.00	.00	
RP-Coun Process	.00	.00	.00	.00	.00	14.21	.00	.00	
RP-Training	.00	.00	.00	.00	.00	14.23	.00	.00	
SJT-Most/Least	.00	.00	7.41	.00	.00	.00	8.03	.00	
Psi (t - values for factor correlations)									
	CT	GP	EA	PD	PF	TC	Written	Ratings	
CT-Job Specific	.00								
GP-General Proficiency	9.10	.00							
EA-Effort	7.56	7.73	.00						
PD-Personal Discipline	-3.18	-1.62	-10.85	.00					
PF-Phys Fitness	.00	.00	13.22	-9.01	.00	.00			
TC-Training/Counseling	5.75	6.22	9.24	-4.19	5.35	.00	.00	.00	
Written	.00	.00	.00	.00	.00	.00	.00	.00	
Ratings	.00	.00	.00	.00	.00	.00	.00	.00	.00

Table 6.5

LISREL Results for Training and Counseling Factor Model: Fitted and Normalized Residuals

Fitted Residuals

	PF Award	PF Art	PF Cour	Grad Dev	PF Phys	HO Gen	HO Job	JK Gen	JK Job	AMB Lead
PFF-Awards	.032									
PFF-Art/Flag	.064	.038								
PFF-Courses	.022	-.012	.043							
Grade Deviation	-.004	-.009	.003	.046						
PFF-Phys Read	-.019	-.024	.032	.011	.081					
HO-General	-.003	-.018	.054	.040	-.007	.036				
HO-Job Specific	-.009	.032	.029	.043	.012	.049	.060			
JK-General	-.046	.026	.060	.040	-.054	.050	.083	.141		
JK-Job Specific	.001	.002	.012	.024	-.065	.011	.038	.076		
AMB-Leadership	.021	-.008	.025	.021	.013	-.003	.014	.040	.060	.051
AMB-Tech Prof	.032	-.009	.033	-.007	.035	.009	.060	.105	.055	.059
AMB-Discipline	-.006	-.013	.029	.008	-.016	.033	.040	.074	.048	.054
AMB-Phys Fitness	-.005	-.020	.070	.019	.040	.047	.076	.083	.042	.057
MOB-MOS Total	.022	.003	.054	.024	.005	.018	.049	.090	.047	.051
RP-Disc Content	.006	.029	-.002	.029	-.014	.011	.004	-.046	-.008	-.040
RP-Disc Process	.018	-.012	.071	-.072	.095	-.036	.002	-.058	.021	-.056
RP-Coun Content	.011	-.016	.007	-.034	.011	.017	-.065	.035	.051	-.030
RP-Coun Process	-.015	-.058	.012	-.053	.015	-.001	-.024	.012	.024	.058
RP-Training	-.006	-.040	.034	.024	.043	.067	.035	.121	.040	.071
SJT-Most/Least	-.027	.048	-.026	.034	-.134	.044	.067	.097	.093	.049

Fitted Residuals (Continued)

	AMB Tech	AMB Disc	AMB Phys	MOB	RPDC	RPDP	RPCC	RPCP	RPT	SJT
AMB-Tech Proficiency	.088									
AMB-Discipline	.067	.078								
AMB-Phys Fitness	.065	.064	.075							
MOB-MOS Total	.076	.059	.062	.075	.124	.123				
RP Disc Content	-.028	-.064	-.069	-.035	-.037	-.015	.084			
RP-Disc Process	-.061	-.053	-.095	-.100	.136	-.030	.096	.146		
RP-Coun Content	-.053	-.039	-.027	-.061	-.036	-.075	-.014	.049	.070	
RP-Coun Process	.078	.115	.076	.041	.166	-.030	.096	.087	.073	.119
RP-Training	.072	.035	.064	.080	-.008	-.075	-.014			
SJT-Most/Least	.067	.087	-.008	.031	.012	.046	.107			

(continued)

Table 6.6

LISREL Results for "Overfit Model": Parameter Estimates

LISREL Estimates (Generalized Least Squares): "Overfit Model"

Lambda Y (Factor loadings)

	CT Job Spec	GP Gen Pref	EA Effort	PD Pers Disc	PF Phys Fit	TC Train	Written	Ratings
PFF-Awards	.00	-.16	.64	.24	.00	.00	.00	.00
PFF-Art/Flag	.00	.00	.00	.41	.00	.00	.00	.00
PFF-Courses	.00	.00	.63	.00	.00	.00	.00	.00
Grade Deviation	.00	.00	.63	.00	.00	.00	.00	.00
PFF-Phys Read	.00	.00	.00	.00	.45	.00	-.19	.00
HO-General	.00	.43	.00	.00	.00	.00	.00	.00
HO-Job Specific	.52	.00	.00	.00	.00	.00	.00	.00
JK-General	.00	.71	.00	.00	.00	.00	.44	.00
JK-Job Specific	.79	.00	.00	.00	.00	.00	.35	.00
AMB-Leadership	.00	-.10	.49	.00	.00	.00	.00	.78
AMB-Tech Prof	.00	.00	.32	.00	.00	.00	.00	.84
AMB-Discipline	.00	.00	.00	-.41	.00	.00	.00	.71
AMB-Phys Fitness	.00	.00	.00	.00	.67	.00	.00	.50
MOB-MOS Total	.00	.00	.29	.00	.00	.00	.00	.75
RP-Disc Content	.00	.00	.00	.00	.00	.49	.00	.00
RP-Disc Process	.00	.00	.00	.00	.00	.34	.00	.00
RP-Coun Content	.00	.00	.00	.00	.00	.67	.00	.00
RP-Coun Process	.00	.00	.00	-.14	.00	.24	.00	.11
RP-Training	.00	.15	.00	.00	.00	.43	.00	.10
SJT-Most/Least	.00	.13	.16	.00	.00	.11	.52	.00

Psi (factor correlations)

	CT	GP	EA	PD	PF	TC	Written	Ratings
CT-Job Specific	1.00							
GP-General Proficiency	.58	1.00						
EA-Effort	.33	.37	1.00					
PD-Personal Discipline	-.11	.00	-.63	1.00				
PF-Phys Fitness	.00	.00	.60	-.64	1.00			
TC-Training/Counseling	.26	.28	.40	-.20	.23	1.00		
Written	.00	.00	.00	.00	.00	.16	1.00	
Ratings	.00	.00	.00	.00	.00	.09	.16	1.00

(Continued)

Table 6.6 (Continued)

LISREL Results for "Overfit Model": Parameter Estimates

LISREL Estimates (Generalized Least Squares): "Overfit Model"

Theta Epsilon (Unique components)

	PF Award	PF Art	PF Cour	Grad Dev	PF Phys	HO Gen	HO Job	JK Gen	JK Job	AMB Lead
PFF Awards	.75									
PFF Art/Flag	.00	.81								
PFF Courses	.00	.00	.58							
Grade Deviation	.00	.00	.00	.56						
PFF Phys Read	.00	.00	.00	.00	.72					
HO General	.00	.00	.00	.00	.00	.78	.69			
HO Job Specific	.00	.00	.00	.00	.00	.00	.00	.29		
JK General	.00	.00	.00	.00	.00	.00	.00	.00	.23	
JK Job Specific	.00	.00	.00	.00	.00	.00	.00	.00	.00	.16
AMB Leadership	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
AMB Tech Prof	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
AMB Discipline	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
AMB Phys Fitness	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
MOB MDS Total	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
RP Disc Content	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
RP Disc Process	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
RP Coun Content	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
RP Coun Process	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
RP Training	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
SJT Most/Least	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Theta Epsilon (Continued)

	AMB Tech	AMB Disc	AMB Phys	MOB	RPDC	RPDP	RPCC	RPCP	RPT	SJT
AMB Tech Proficiency	.16									
AMB Discipline	.00	.30								
AMB Phys Fitness	.00	.00	.25							
MOB MDS Total	.00	.00	.00	.31	.73	.83				
RP Disc Content	.00	.00	.00	.00	.21	.45				
RP Disc Process	.00	.00	.00	.00	.00	.00	.51			
RP Coun Content	.00	.00	.00	.00	.00	.00	.22	.87	.73	
RP Coun Process	.00	.00	.00	.00	.04	.00	.16	.16	.00	
RP Training	.00	.00	.00	.00	.00	.00	.00	.00	.00	.59
SJT Most/Least	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Table 6.7**LISREL Results for "Overfit Model": Fit Statistics**

Measures of Goodness-of-Fit for the Whole Model:

Chi-Square with 132 Degrees of Freedom	171.03 (Prob. Level = 0.013)
Goodness-of-Fit Index	.983
Adjusted Goodness-of-Fit Index	.973
Root mean square residual	.030

Subsample Differences

Available sample sizes did not permit separate analyses for different race or gender groups or by MOS. We were, however, concerned about differences in supervisory experience and the possible effect that these differences might have on the identification of a "supervision" factor. We used available indicators to separate the sample into soldiers with more or less supervisory responsibility, using items from the background questionnaire and the willingness of supervisors to provide ratings on supervisory dimensions. Table 6.10 reports the means for the various criterion measures for the overall sample, and for the two subsamples. Correlations among the criterion measures for the subsamples are reported in Tables 6.11 and 6.12.

In general, differences between the two subsamples were not very significant. At one point, one of the alternative factor models was fit to each group separately, and the results were compared to the fit when the parameters were all constrained to be identical for the two groups. The difference, a chi-square of 94.1 with 96 degrees of freedom, was not at all significant. Further attempts to differentiate these two groups were therefore abandoned.

FINAL SCORES

Scores were computed for each of the six substantive performance factors in the second-tour model. Figure 6.1 shows the assignment of specific variables to overall performance factors. In computing these scores, variables were standardized, combined within method (where appropriate), and then restandardized to a mean of 50 and a standard deviation of 10. Corresponding scores for different methods were then combined and the resulting sums were standardized to a mean of 100 and a standard deviation of 10. This process gives equal weight to each measurement method, minimizing potential measurement bias for all factors except the Training and Counseling factor (which is covered by only one method). Table 6.13 shows the correlation between the resulting scores and the component variables with asterisks used to indicate the variables used in computing each factor score.

Table 6.8

LISREL Results for "Overfit Model": t-ValuesLambda Y (t-values for factor loadings)

	CT Job Spec	GP Gen Pref	EA Effort	PD Pers Disc	PF Phys Fit	TC Train	Written	Ratings
PFF-Awards	.00	-2.79	7.44	2.74	.00	.00	.00	.00
PFF-Art/Flag	.00	.00	.00	8.72	.00	.00	.00	.00
PFF-Courses	.00	.00	17.74	.00	.00	.00	.00	.00
Grade Deviation	.00	.00	17.82	.00	.00	.00	.00	.00
PFF-Phys Read	.00	.00	.00	.00	11.12	.00	-4.03	.00
HO-General	.00	10.25	.00	.00	.00	.00	.00	.00
HO-Job Specific	11.80	.00	.00	.00	.00	.00	.00	.00
JK-General	.00	13.81	.00	.00	.00	.00	7.36	.00
JK-Job Specific	14.56	.00	.00	.00	.00	.00	6.89	.00
AMB-Leadership	.00	-4.09	13.87	.00	.00	.00	.00	31.18
AMB-Tech Prof	.00	.00	9.45	.00	.00	.00	.00	33.61
AMB-Discipline	.00	.00	.00	-9.98	.00	.00	.00	25.59
AMB-Phys Fitness	.00	.00	.00	.00	.00	.00	.00	17.69
MOB-WDS Total	.00	.00	8.42	.00	14.83	.00	.00	28.41
RP-Disc Content	.00	.00	.00	.00	.00	10.16	.00	.00
RP-Disc Process	.00	.00	.00	.00	.00	6.05	.00	.00
RP-Coun Content	.00	.00	.00	.00	.00	12.10	.00	.00
RP-Coun Process	.00	.00	.00	-3.11	.00	4.16	.00	3.49
RP-Training	.00	3.43	.00	.00	.00	8.28	.00	3.07
SJT-Most/Least	.00	2.23	3.80	.00	.00	2.49	7.74	.00

Psi (t-values for factor correlations)

	CT	GP	EA	PD	PF	TC	Written	Ratings
CT-Job Specific	.00							
GP-General Proficiency	11.96	.00						
EA-Effort	8.13	8.53	.00					
PD-Personal Discipline	-2.23	.00	-11.57	.00				
PF-Phys Fitness	.00	.00	13.97	-10.36	.00			
TC-Training/Counseling	5.24	4.98	8.35	-3.01	4.44	.00		
Written	.00	.00	.00	.00	.00	.00	.00	
Ratings	.00	.00	.00	.00	.00	-2.02	3.05	.00

Theta Epsilon (covariance among uniquenesses)

	RPDC	RPDP	RPCC	RPCP	RPT
RP-Disc Content	14.59				
RP-Disc Process	6.15	16.68			
RP-Coun Content	.00	1.09	7.44		
RP-Coun Process	.00	3.37	5.25	19.36	
RP-Training	1.05	.00	.00	5.16	16.01

Table 6.9

LISREL Results for "Overfit Model": Fitted and Normalized Residuals

Fitted Residuals

	PF Award	PF Art	PF Cour	Grad Dev	PF Phys	HO Gen	HO Job	JK Gen	JK Job	AMB Lead
PFF-Awards	.022									
PFF-Art/Flag	.020	.026								
PFF-Courses	.022	-.011	.033							
Grade Deviation	.000	-.010	.003	.045						
PFF-Phys Read	-.010	-.011	.025	.005	.043					
HO-General	.041	-.039	.067	.054	-.007	.033				
HO-Job Specific	.010	.020	.020	.034	.012	.044	.041			
JK-General	-.011	.001	.029	.010	.027	-.006	.015	.020		
JK-Job Specific	.033	-.019	.001	.014	.000	.007	.007	.008	.025	
AMB-Leadership	.016	.001	.026	.025	.023	.044	.031	.023	.008	.025
AMB-Tech Prof	.034	-.009	.034	-.004	.057	.016	.056	.030	.003	.024
AMB-Discipline	.041	.004	.033	.013	-.005	.055	.054	.049	.030	.020
AMB-Phys Fitness	.007	.001	.056	.008	.038	.047	.076	.047	.014	.025
MOB-MOS Total	.021	.005	.049	.021	.024	.023	.044	.021	-.003	.021
RP-Disc Content	.017	.014	.007	.039	-.004	.038	.006	-.042	-.002	.009
RP-Disc Process	.041	-.031	-.044	-.044	-.079	-.003	.015	-.039	.043	-.006
RP-Coun Content	.013	-.028	-.001	-.040	.016	.042	-.072	.026	.044	.023
RP-Coun Process	.007	-.033	-.006	-.045	.010	.049	-.001	.050	.055	-.003
RP-Training	-.004	-.055	.013	.004	.057	.031	-.004	.017	-.023	.029
SJT-Most/Least	-.008	.030	-.018	.044	-.024	-.004	.026	.015	.033	.011

Fitted Residuals (Continued)

	AMB Tech	AMB Disc	AMB Phys	MOB	RPDC	RPDP	RPCC	RPCP	RPT	SJT
AMB-Tech Proficiency	.035									
AMB-Discipline	.022	.027								
AMB-Phys Fitness	.037	.028	.047							
MOB-MOS Total	.026	.020	.036	.033						
RP-Disc Content	.015	-.015	-.032	.002	.026	.054				
RP-Disc Process	-.020	-.011	-.055	-.065	.005	.017	.046			
RP-Coun Content	-.004	.019	.011	-.018	-.006	.038	.031	.038	.026	
RP-Coun Process	.010	.029	.018	-.022	-.021	.013	.006	.012	-.008	
RP-Training	.012	.009	.049	.024	-.003	-.001	.062	.026		
SJT-Most/Least	.010	.055	-.024	-.022	-.037	.021	.031	.062		.051

(Continued)

Table 6.9 (Continued)

LISREL Results for "Overfit Model": Fitted and Normalized Residuals

<u>Normalized Residuals</u>										
	PF Award	PF Art	PF Cour	Grad Dev	PF Phys	MO Gen	MO Job	JK Gen	JK Job	AMB Lead
PFF-Awards	.511									
PFF-Art/Flag	.660	.591								
PFF-Courses	.706	-.359	.755							
Grade Deviation	-.011	-.312	.095							
PFF-Phys Read	-.315	-.350	.797							
MO-General	1.333	-1.263	2.184		1.006					
MO-Job Specific	.341	.648	.644		-.228	.776	.965			
JK-General	-.341	.026	.924	1.135	.408	1.440	.478	.451		
JK-Job Specific	1.065	-.612	.031	.320	.879	-.177	.223	.227	.571	
AMB-Leadership	.525	.034	.818	.790	.752	1.433	1.022	.740	-.264	.572
AMB-Tech Prof	1.112	-.296	1.095	-.133	1.889	.536	1.833	.967	.087	.611
AMB-Discipline	1.338	.125	1.053	.422	-.166	1.803	1.776	1.606	.979	.524
AMB-Phys Fitness	.218	.043	1.802	.262	1.204	1.562	2.525	1.537	.452	.699
MOB-MOS Total	.670	.170	1.588	.677	.780	.760	1.457	.670	-.081	.549
RP-Disc Content	.544	.461	.212	1.264	-.143	1.254	.211	-1.368	-.072	.308
RP-Disc Process	1.340	-1.024	-1.466	-1.455	-2.625	-.113	.493	-1.271	1.403	-.203
RP-Coun Content	.420	-.934	-.025	-1.296	.530	1.374	-2.374	.838	1.440	.753
RP-Coun Process	.226	-1.070	-.183	-1.474	.325	1.610	-.032	1.624	1.811	-.106
RP-Training	-.131	-1.789	.431	.143	1.881	.991	-.137	.526	-.735	.925
SJT-Most/Least	-.276	.999	-.590	1.431	-.785	-.132	.847	.469	1.053	.372
<u>Normalized Residuals (Continued)</u>										
	AMB Tech	AMB Disc	AMB Phys	MOB	RPDC	RPDP	RPCC	RPCC	RPT	SJT
AMB-Tech Proficiency	.815									
AMB-Discipline	.593	.612								
AMB-Phys Fitness	1.051	.819	1.116							
MOS-MOS Total	.680	.544	1.056	.776						
RP-Disc Content	.502	-.497	-1.039	.073	.596					
RP-Disc Process	-.663	-.352	-1.827	-2.149	.169	1.294	1.093			
RP-Coun Content	-.144	.628	.360	-.609	-.187	.539	.940	.883		
RP-Coun Process	.324	.951	.606	-.721	-.685	.424	.387	.387	.589	
RP-Training	.384	.307	1.622	.778	-.100	-.028	.195	.387	-.271	
SJT-Most/Least	.327	1.806	-.803	-.728	-1.228	.689	1.037	2.069		1.209

Table 6.10

Means of the CVII Summary Measures for All Soldiers* and for Soldiers With and Without Supervisory Experience

Criterion Variable	Total Sample (N = 1,009)		Supervisors (N = 542)		Nonsupervisors (N = 467)	
	Mean	SD	Mean	SD	Mean	SD
<u>Personnel File Folder</u>						
1 PFF: Awards, memos, certificates	10.58	5.70	11.51	5.68	9.50	5.53
2 PFF: No. Flags & Articles at/above E-4	.43	.89	.35	.79	.51	1.00
3 PFF: No. training courses taken	1.37	1.03	1.66	1.01	1.03	.94
4 PFF: Grade deviation	99.97	8.27	102.11	8.21	97.48	7.63
5 PFF: SQT Score	79.41	9.31	80.53	8.71	78.11	9.80
6 PFF: Physical Readiness Score	249.88	30.57	253.24	29.63	245.99	31.20
7 PFF: M16/M19 Qualification	2.53	.67	2.66	.60	2.37	.71
<u>Hands-On Tests</u>						
8 H0 Basic/Tech: Basic Soldiering	.73	.15	.75	.14	.71	.15
9 H0 Basic/Tech: MOS Specific	.69	.20	.72	.19	.65	.20
<u>Job Knowledge Tests</u>						
10 JK Basic/Tech: Basic Soldiering	.65	.12	.67	.12	.63	.11
11 JK Basic/Tech: MOS Specific	.65	.14	.67	.14	.63	.13
<u>Supervisor Ratings</u>						
12 AWB: Lead/Supv	4.49	1.06	4.73	1.00	4.21	1.06
13 AWB: Tech Prof/Skill	5.03	1.09	5.20	1.04	4.84	1.10
14 AWB: Personal Discipline	5.16	1.11	5.31	1.03	4.99	1.19
15 AWB: Phys Fit/Mil Bearing	5.19	1.19	5.38	1.11	4.97	1.23
16 MOB: MOS Total	5.19	.97	5.33	.88	5.04	1.04
<u>Simulation Exercises</u>						
17 Role: Disciplinary Content	2.01	.47	2.04	.49	1.97	.46
18 Role: Disciplinary Process	2.45	.48	2.47	.47	2.42	.49
19 Role: Counseling Content	1.81	.46	1.87	.47	1.74	.45
20 Role: Counseling Process	2.66	.33	2.68	.31	2.63	.36
21 Role: Training	2.38	.48	2.41	.45	2.34	.50
<u>Situational Judgment Test</u>						
22 SJT Most/Least	1.37	.60	1.45	.57	1.29	.62

*After minimal imputations of missing data.

Table 6.11

Correlations Among the CVII Summary Measures Based on Soldiers With Supervisory Experience

Criterion Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
	Awd	Art	TrC	Grd	SQT	PhR	M16	HOG	HOS	JKG	JKS	R-L	R-T	R-D	R-F	MOB	RDC	RDP	RCC	RCP	RT	RJT
<u>Personnel File Folder</u>																						
1 PFF-Awards	xx	-02	21	17	15	04	09	05	03	03	10	17	14	07	09	08	12	10	15	09	06	09
2 PFF-Art/Flag	-02	xx	-15	-12	-00	-17	-02	-03	02	-02	-02	-11	-07	-12	-14	-05	01	-01	02	-09	-08	03
3 PFF-Courses	21	-15	xx	32	16	17	15	16	10	13	10	30	21	16	30	23	11	-02	13	08	11	11
4 Grade Deviation	17	-12	32	xx	14	15	07	12	13	15	15	28	17	14	23	19	19	-03	06	01	10	19
5 PFF-SQT	15	-00	16	14	xx	01	09	11	22	21	22	17	18	11	08	20	13	-03	-00	03	12	15
6 PFF-Phys Read	04	-17	17	15	01	xx	03	05	-01	-06	-07	08	09	03	31	05	04	01	05	06	15	-08
7 PFF-M16/19	09	-02	15	07	09	03	xx	01	03	04	00	08	08	-02	07	15	-00	-10	-04	-02	08	03
<u>Hands-On Tests</u>																						
8 HO-General	05	-03	16	12	11	05	01	xx	17	26	22	09	08	09	07	05	14	01	09	05	11	12
9 HO-Job Specific	03	02	10	13	22	-01	03	17	xx	20	44	04	09	06	08	10	07	04	00	05	14	08
<u>Job Knowledge Tests</u>																						
10 JK-General	03	-02	13	15	21	-06	04	26	20	xx	47	09	14	06	07	11	02	-03	14	07	19	40
11 JK-Job Specific	10	-02	10	15	22	-07	00	22	44	47	xx	07	09	06	01	07	08	06	19	09	11	28
<u>Supervisor Ratings</u>																						
12 AWB-Leadership	17	-11	30	28	17	08	08	09	04	09	07	xx	83	66	57	73	04	02	04	14	10	09
13 AWB-Tech Prof	14	-07	21	17	18	09	08	08	09	14	09	83	xx	69	55	73	02	-03	-01	12	09	08
14 AWB-Discipline	07	-12	16	14	11	03	-02	09	06	06	06	66	69	xx	54	60	-01	-04	-04	14	03	09
15 AWB-Phys Fitness	09	-14	30	23	08	31	07	07	08	07	01	57	55	54	xx	48	03	-00	00	14	11	07
16 MOB-MOS Total	08	-05	23	19	20	05	15	05	10	11	07	73	73	60	48	xx	-00	-06	-03	09	06	02
<u>Simulation Exercises</u>																						
17 RP-Disc Content	12	01	11	19	13	04	-00	14	07	02	08	04	02	-01	03	-00	xx	37	31	05	24	09
18 RP-Disc Process	10	-01	-02	-03	-03	01	-10	01	04	-03	06	02	-03	-04	-00	-06	37	xx	24	11	14	03
19 RP-Coun Content	15	02	13	06	-00	05	-04	09	00	14	19	04	-01	-04	00	-03	31	24	xx	40	30	16
20 RP-Coun Process	09	-09	08	01	03	06	-02	05	05	07	09	14	12	14	14	09	05	11	40	xx	26	10
21 RP-Training	06	-08	11	10	12	15	08	11	14	19	11	10	09	03	11	06	24	14	30	26	xx	09
<u>Situational Judgment Test</u>																						
22 SJT-Most/Least	09	03	11	19	15	-08	03	12	08	40	28	09	08	09	07	02	09	03	16	10	09	xx

Note. N = 542. Decimals have been omitted.

Table 6.12

Correlations Among the CVII Summary Measures Based on Soldiers Without Supervisory Experience

Criterion Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
	Awd	Art	TrC	Grd	SQT	PhR	M16	HOG	HOS	JKG	JKS	R-L	R-T	R-D	R-F	MOB	RDC	RDP	RCC	RCP	RT	SJT
Personnel File Folder																						
1 PFF-Awards	xx	-04	32	32	21	12	12	07	03	00	05	20	17	10	18	17	05	08	04	02	07	05
2 PFF-Art/Flag	-04	xx	-16	-19	-03	-08	-00	-03	01	05	-06	-11	-09	-18	-18	-06	-05	-09	-16	-11	-09	-04
3 PFF-Courses	32	-16	xx	37	21	17	13	12	05	17	13	19	18	16	24	17	11	09	12	10	17	10
4 Grade Deviation	32	-19	37	xx	11	15	10	14	06	10	12	23	15	16	22	16	10	10	13	10	17	16
5 PFF-SQT	21	-03	21	11	xx	02	08	18	15	28	26	14	19	09	03	21	09	10	09	11	12	19
6 PFF-Phys Read	12	-08	17	15	02	xx	11	-10	-00	-09	-10	13	11	12	32	07	04	-11	08	07	03	-08
7 PFF-M16/19	12	-00	13	10	08	11	xx	16	11	05	-05	07	08	01	09	07	-08	-20	-09	-02	07	-02
Hands-On Tests																						
8 HO-General	07	-03	12	14	18	-10	16	xx	15	31	15	02	01	-01	-01	06	04	06	13	09	17	04
9 HO-Job Specific	03	01	05	06	15	-00	11	15	xx	21	36	05	08	05	02	04	06	07	-02	01	04	09
Job Knowledge Tests																						
10 JK-General	00	05	17	10	28	-09	05	31	21	xx	47	11	16	09	04	14	07	08	13	12	22	37
11 JK-Job Specific	05	-06	13	12	26	-10	-05	15	36	47	xx	12	13	12	02	12	10	16	13	14	15	38
Supervisor Ratings																						
12 AWB-Leadership	20	-11	19	23	14	13	07	02	05	11	12	xx	80	72	62	74	04	01	06	13	19	20
13 AWB-Tech Prof	17	-09	18	15	19	11	08	01	08	16	13	80	xx	70	60	76	04	01	02	15	16	19
14 AWB-Discipline	10	-18	16	16	09	12	01	-01	05	09	12	72	70	xx	57	64	-02	01	06	18	12	19
15 AWB-Phys Fitness	18	-18	24	22	03	32	09	-01	02	04	02	62	60	57	xx	55	-02	-05	12	16	16	08
16 MOB-MOS Total	17	-06	17	16	21	07	07	06	04	14	12	74	76	64	55	xx	03	-05	01	08	18	15
Simulation Exercises																						
17 RP-Disc Content	05	-05	11	10	09	04	-08	04	06	07	10	04	04	-02	-02	03	xx	41	32	15	28	02
18 RP-Disc Process	08	-09	09	10	10	-11	-20	06	07	08	16	01	01	01	-05	-05	41	xx	35	29	17	15
19 RP-Coun Content	04	-16	12	13	09	08	-09	13	-02	13	13	06	02	06	12	01	32	35	xx	44	32	16
20 RP-Coun Process	02	-11	10	10	11	07	-02	09	01	12	14	13	15	18	16	08	15	29	44	xx	33	16
21 RP-Training	07	-09	17	17	12	03	07	17	04	22	15	19	16	12	16	18	28	17	32	33	xx	14
Situational Judgment Test																						
22 SJT-Most/Least	05	-04	10	16	19	-08	-02	04	09	37	38	20	19	19	08	15	02	15	16	16	14	xx

Note. N = 647. Decimals have been omitted.

Latent Variables in the CVII Performance Model:

- **Core Technical Proficiency (CT)**
 - Job-Specific Hands-On
 - Job-Specific Knowledge
- **General Soldiering Proficiency (GS)**
 - Common Hands-On
 - Common Job Knowledge
- **Effort and Leadership (EL)**
 - Awards and Certificates
 - Training Courses
 - Grade Deviation Score
 - Army-Wide BARS Leadership Rating
 - Army-Wide BARS Technical Rating
 - MOS BARS Average Rating
 - Situational Judgment Test
- **Personal Discipline (PD)**
 - Articles 15, Flag Actions (reversed)
 - Army-Wide BARS Discipline Rating
- **Physical Fitness/Military Bearing (PF)**
 - Physical Readiness Score
 - Army-Wide BARS Fitness/Bearing Rating
- **Training and Counseling Subordinates (TC)**
 - Simulation Exercise - Personal Counseling Content
 - Simulation Exercise - Personal Counseling Process
 - Simulation Exercise - Disciplining Content
 - Simulation Exercise - Disciplining Process
 - Simulation Exercise - Training
- **Written Methods (WM)**
 - Job-Specific Knowledge
 - Common Soldiering Knowledge
 - Situational Judgment Test
- **Ratings Methods (RM)**
 - Four Army-Wide BARS Dimensions
 - MOS BARS Average

Figure 6.1. Relationship of specific variable to overall factors in the CVII performance model.

Table 6.13

Correlation of Specific Measures With Provisional Performance Scores for CVII^a

Variable	Core Technical	General Soldiering	Effort/Leadership	Personal Discipline	Physical Fitness	Training/Counseling
Awards & Certs	.15	.11	.53*	.10	.15	.14
Articles 15	-.05	-.00	-.13	-.76*	-.18	-.09
Training Courses	.19	.21	.40*	.23	.30	.18
Grade Deviation	.20	.20	.61*	.22	.26	.16
Self-Report	.38	.34	.28	.10	.06	.13
Physical Readiness	.01	-.05	.12	.14	.81*	.09
M16	.07	.10	.16	.04	.12	-.02
HO-Common	.26	.82*	.19	.05	.02	.15
HO-Specific	.83*	.25	.18	.08	.07	.13
JK-Common	.45	.82*	.38	.07	.02	.20
JK-Specific	.83*	.45	.34	.12	.04	.16
AWB-Leadership	.17	.17	.69*	.54	.45	.13
AWB-Technical	.19	.18	.64*	.53	.44	.11
AWB-Discipline	.13	.10	.51	.76*	.40	.06
AWB-Fitness	.08	.09	.48	.48	.81*	.11
MOB-Job Specific	.19	.18	.59*	.47	.37	.07
RP-Disc/Content	.10	.09	.12	-.00	.05	.65*
RP-Disc/Process	.09	.04	.07	.02	-.02	.63*
RP-Pers/Content	.08	.16	.17	.06	.10	.69*
RP-Pers/Process	.09	.12	.17	.17	.14	.54*
RP-Training	.18	.23	.20	.11	.14	.64*
SJT	.24	.32	.66*	.11	.04	.16
CT Construct	1.00	.43	.32	.12	.05	.17
GS Construct	.43	1.00	.35	.07	.02	.21
EL Construct	.32	.35	1.00	.42	.37	.23
PD Construct	.12	.07	.42	1.00	.38	.10
PF Construct	.05	.02	.37	.38	1.00	.12
TC Construct	.17	.21	.23	.10	.12	1.00

Note. N = 1006-1009. Standardized by MOS, after pseudo-imputations.
Mean = 100; standard deviation = 19.92.

^a * indicates the variables used in computing each factor score.

IMPLICATIONS

Given the limits of the criterion data and our ability to theorize, the solution of six substantive factors plus two method factors represents our best current portrayal of the latent structure of second-tour performance. It extends our knowledge of performance in a number of ways but also carries some important limitations. The major points can be summarized as follows.

The second-tour job analyses produced something of a surprise when they described the significant role of one-on-one counseling type tasks in the job of second-tour NCO. Consequently, a lot of attention was devoted in a short space of time to developing simulation procedures to measure this part of the performance domain. The course of events did not permit extensive development of multiple measures of counseling performance. Consequently, the six-factor model confounds substantive variance and potential method variance more than we would like. However, it is noteworthy that a model which hypothesized a role-play simulation methods factor and then attempted to assign the residual variance from the simulations to other substantive factors did not fit nearly so well. There is something to counseling/training performance as a distinct latent component.

The evidence for the predicted correspondence between the first-tour model and the second-tour model was gratifying. Virtually all the basic variables looked as they should, given expectations based on CVI results. One disappointment was the lack of evidence for more specific leadership components within the general leadership factor. This could represent either a failure of the current rating scales to reflect latent variables that do exist or the fact that the CVII sample is still quite young and has had relatively few opportunities to exhibit leadership performance. Most likely it is some of both.

A second-tour criterion measure of special interest is the Situational Judgment Test (SJT). It is a paper-and-pencil measure intended to measure the quality of supervisory/leadership judgment in relatively unstructured problem situations. Perhaps because of its common method, the SJT's highest zero-order correlations are with the two scores from the job knowledge tests. However, its secondary correlations are with measures that compose the effort/leadership factor, and when the paper-and-pencil methods factor is extracted from all such criterion measures, the SJT clearly fits with the general leadership factor. Given that its reliability is high, its correlations with other variables are relatively low, and the relevant variance is most closely associated with the effort/leadership construct, it may be the case that the SJT provides a significant amount of genuinely unique variance relative to leadership performance. Further analysis of selected SJT residual scores should shed more light on this question.

Finally, the relevant variance in the "grade deviation" scores seems to change when moving from first tour to second tour. The score is an indicator of an individual's relative rate of promotion. During the first tour it seemed most dependent on whether the soldier lacks personal discipline and gets into trouble, while in the second tour it relates more to recognition and achievement (or lack of it). That is, for people in their second tour, a relatively high promotion rate is due to positive achievement rather than simply the avoidance of trouble.

In sum, the second-tour performance model is quite consistent with the first-tour model, in terms of both its similarities and its differences. The six-factor model provides a good fit with the data and suggests some interesting empirical relationships to investigate further. Additional construct validation is warranted and would be greatly facilitated by additional redundancy and multiple measurement methods for the supervisor/leadership domain. The CVII samples will provide an opportunity to follow-up on these findings and to extend them.

Chapter 7

Future Career Force Project Plans

Two central activities will dominate work on the Career Force Project during the second contract year. One deals with the LVII data collection and the other with the planned analyses of the data from the Longitudinal Validation (LVI), end-of-training (EOT), and second-tour Concurrent Validation (CVII) samples. This chapter outlines the near-term plans for each of these major activities.

LVII DATA COLLECTION

One of the central purposes of the Career Force Project is to continue to track the LVI sample from Project A and to assess the reenlistees on the second-tour measures of job performance as they begin to take on NCO supervisory and leadership responsibilities. The measurement of NCO performance will be used to examine how the ASVAB and the Project A Predictor Battery, in conjunction with first-tour performance measures, are related to NCO performance. The potential benefits to the Army from improvements in NCO selection and classification are great, as are the potential contributions to our general knowledge about the long-term effects of different selection and classification procedures.

The LVII data collection is scheduled to occur between May 1991 and February 1992. The goal is to obtain data from a minimum of 150 soldiers from each of the following MOS: 11B, 13B, 19K, 31C, 63B, 71L, 88M, 91A, and 95B. These are the MOS for which we have previously administered both written and hands-on tests. In Project A, they became identified as the "Batch A" MOS.

Anticipated Problems

Relatively speaking, this data collection will be very resource-intensive in terms of the number of soldiers who must be tested and the equipment, facilities, and military staffing that will be needed to support the testing effort. Furthermore, the design of the research requires us to be unusually selective in the specific soldiers to be tested in that they must already be members of the Project A Longitudinal Validation sample.

A variety of problems can be anticipated with a data collection effort of this type. Project A testing, particularly in the LVI/CVII data collection, provided first-hand experience with some of these problems, and it is hoped that this experience will lead to improvements in the LVII data collection. In addition to facing some unique problems stemming from the nature of the inquiry, we can expect some of the usual data collection problems to be exacerbated by the current state of uncertainty in the Army resulting from the deployment of soldiers to the Middle East and projected personnel downsizing efforts.

Advance Preparation of Research Support Requests. An established, formal procedure must be followed when requesting troops and other support required to conduct data collections in the Army. The process is designed to give installations providing research support considerable advance notice about details of the requirements. Thus, the Research Support Requests (RSRs) for the LVII data collection have to be submitted a full year before the data collection is scheduled to begin.

The RSRs must include the number of soldiers from each MOS who are being requested from each installation. Ordinarily, this would not be a significant problem since soldiers are often requested only with regard to their MOS or paygrade. Because this is a longitudinal data collection, however, the soldiers who will be requested for participation in LVII will be requested by name. We refer to those soldiers who have participated in earlier Project A testing, and who are therefore eligible for testing in 1991, as our "target" soldiers.

It is not easy to determine the current location of individual soldiers because the records to which we have access are always somewhat out of date by the time we get them. It is considerably more difficult to anticipate where individual soldiers will be one to two years hence. Accordingly, preparing detailed RSRs which correspond to the actual location of target soldiers at the time of testing is a gamble at best.

Small Numbers of Soldiers Eligible for Testing. Although several thousand soldiers were tested with the Project A predictors and first-tour criterion measures, a majority of these individuals will not be available for testing in 1991. The primary reason will be the fact that many soldiers do not reenlist for a second term of service. Historically, reenlistment rates for the MOS to be tested have ranged from 31 percent (31C) to 48 percent (71L). Based on these figures, we can expect fewer than half of the target soldiers to reenlist. This figure could dramatically decrease in the downsizing of the military services. Downsizing could also result in forced early separations from the Army.

Another reason for decreases in the number of soldiers eligible for testing is that some soldiers will change their MOS. Many soldiers do this when they reenlist. Moreover, for at least one MOS to be tested in 1991, transitions to one of two new MOS will be the result of equipment changes. Specifically, the equipment used by Single Channel Radio Operators (31C) is being phased out. As 31C soldiers are trained to use the new equipment, their MOS will be changed accordingly to either 31D or 31F. Once soldiers have converted to the new equipment, none of the MOS-specific criterion measures will be appropriate for them.

Difficulty Getting Soldiers to Testing. Getting soldiers to be present for testing is invariably a struggle. Factors that interfere with this goal include (a) training or alert status which typically makes soldiers off-limits for testing, (b) soldiers going on leave during the testing period, and (c) lack of planning and/or cooperation from participating installations. Again, the problem is especially difficult when only name-requested soldiers are eligible for testing because substitutes cannot be used for named soldiers who cannot make it to testing.

It is feasible to conduct testing at only a limited number of sites, and soldiers must be able to get to the test sites without traveling long distances. Thus, many target soldiers will not be tested because they will not be close enough to any of the test sites.

Research Support Requests

Given our concerns about being able to test sufficient numbers of target soldiers in LVII, the RSRs were prepared to accommodate a number of contingencies (e.g., soldiers changing locations, some commands or installations not complying with the requests). The result was that, across installations and commands, the RSRs request considerably more soldiers than are needed to meet the sample size goals. This was done with the understanding that the project would cease testing once sufficient data had been collected. The contingencies built into the RSRs are described below. Note that some of them may be eliminated when the RSRs are updated 6 months before data collection begins.

During Project A, criterion data were collected at 13 sites within the continental United States and at several sites located in West Germany. Anticipating that some of these locations would not be able to participate in the LVII data collection, we prepared RSRs for a number of locations that had not previously participated in Project A testing. These "backup" locations include Alaska, Hawaii, and Korea.

Rather than test sites being provided with paper rosters of soldiers we are requesting to test at their location (as had been done in LVI/CVII), they will be provided with a computer file containing the Social Security Numbers (SSNs) of all soldiers eligible for testing. By matching the SSNs on this file with its own computerized personnel records, each site will be able to generate a list of soldiers we want to test who are currently at its location. This will allow the sites to more finely tune their tasking and scheduling activities so that as many target soldiers as possible are reached.

Early in FY90, it was clear that the number of soldiers still in the Army who had both Project A predictor and first-tour criterion data was relatively small. Although we would have preferred to restrict the data collection to soldiers who had both Project A predictor and first-tour criterion data, it was decided at that point that soldiers would be considered eligible for participation in LVII if they had one or the other.

The RSRs developed in the spring of 1990 include requests for soldiers in two MOS, Combat Engineers (12B) and Food Service Specialists (94B), which were in Project A's "Batch Z" test group. This was done to provide a contingency if not enough criterion data could be collected to support validation analyses from the original nine "Batch A" MOS. Unfortunately, no MOS-specific second-tour (or first-tour) criterion measures are available for MOS 12B and 94B. Thus, data from these MOS would increase only the overall sample size for validation analyses related to Army-wide criteria.

A major constraint in the criterion data collection will be the administration of hands-on tests. Administering hands-on tests requires considerably more coordination than written test administration because of the significant equipment, test site, personnel, and scheduling requirements. Without hands-on testing, the probability of testing sufficient numbers of

target soldiers can be greatly improved. Thus, if it appears to be necessary to meet sample size goals, decision rules will be established for eliminating hands-on testing for some MOS. Because 31C is a small-density MOS that is steadily getting smaller, and because the soldiers are particularly widely dispersed throughout the world, the decision has already been made not to administer hands-on tests to soldiers in this MOS.

Revision of Performance Measures

Almost all of the second-tour criterion measures were administered to soldiers in CVII. Minor changes arising from analysis of the CVII data will be made to some measures. For the most part, however, efforts will focus on reviewing the measures to determine whether they are consistent with current doctrine, procedures, and equipment.

The review will be a three-step process. First, project staff will review the latest editions of Soldier Manuals and other relevant procedural documentation to determine what changes might be necessary to update the tests. Second, project staff will review the measures with senior NCOs from each of the tested MOS. This review will include a walk-through of the hands-on tests using actual equipment. Based on the information gathered in these first two steps, the measures will be revised as necessary. Finally, the revised tests will be submitted to each MOS proponent for formal approval.

The only major developmental work in preparation for LVII will be preparation of second-tour job knowledge and hands-on tests for M1 Armor Crewmen (19K). Over the last several years, the 19E MOS (M60 Armor Crewmen) in the original Batch A group has been largely replaced with the 19K MOS. In light of this transition, the 19K was added to the project just prior to the LVI data collection. Since then, the 19E MOS has become too small to retain. Rather than lose the representation of an armor crewmember MOS in LVII, a set of job knowledge and hands-on tests suitable for second-tour 19K soldiers will be constructed.

DATA ANALYSIS PLANS

The future analyses of Career Force Project data can be divided generally into near-term and longer term objectives. The near-term goals encompass those analyses that we plan to accomplish during the second year of the project. The results they generate will be summarized in the annual report for the second year.

Near-Term Analyses

The near-term analyses fall into three categories: (a) Further confirmation and analyses of the CVII and LVI performance models, (b) basic validation analyses for LVI, EOT, and CVII, and (c) the initial validity generalization analyses.

The CVII Performance Model

Chapter 6 presented data to support a model of six substantive factors (plus two methods factors) for second-tour job performance. Further analyses will be carried out to more fully describe the role of the new second-tour measures in assessing the various second-tour performance components. For example, alternative ways of accounting for the variance in the paper-and-pencil method in the SJT when actual criteria scores are generated will be investigated.

Additional analyses will also attempt to specify the hierarchical properties of the CVII model. That is, are there higher order factors that can be established? At how many levels?

The LVI Performance Model

Before the LVI validation of the Experimental Predictor Battery can begin, the first-tour performance factor scores must be generated. As in CVI, these scores will serve as the first-tour criterion measures. The general procedure will be to use the CVI five-factor model as the target against which to conduct a confirmatory analysis using LVI sample data. To increase the yield of information, the project staff will generate a set of alternative a priori models that will be compared to the target for goodness of fit.

The confirmatory analyses will be run twice, corresponding to two different ways of scoring the Hands-On and Job Knowledge measures. In the CVI analyses six subscores were used for each (i.e., the CVBITS category scores). Analyses of the LVI sample data suggest that two subscores for each measure (MOS general and MOS specific) are more parsimonious and do not result in a loss of information. Consequently, the confirmatory analyses will be carried out on both the six-subscore and two-subscore arrangements for Hands-On and Job Knowledge data.

Finally, the hypothesis that one model will fit the data from each individual MOS will also be tested.

Basic Validation Analyses

By basic validation we mean the calculation of criterion-related validity estimates for the relevant predictor set against each criterion factor within each major sample. These analyses will be carried out for the End-of-Training sample (EOT), the Longitudinal Validation sample (LVI), and, to a limited extent, the second-tour Concurrent Validation sample (CVII).

In general, for each sample, the first step will be to estimate validities for each predictor domain (e.g., ASVAB, spatial, perceptual/psychomotor, temperament, interests) against each performance factor. The next step will obtain the incremental validities of the Experimental Battery predictors over the ASVAB for each performance factor. Finally, the full battery will be subjected to a hierarchical analysis for purposes of identifying the optimal battery. These analyses are analogous to what was done in the Concurrent Validation, as reported by McHenry, et al. (1990).

Insofar as possible, the above steps will also be carried out for two sets of a priori predictor weights. One set will consist of the predictor weights obtained in the Concurrent Validation. The second set will be the judgment-based weights identified in the Synthetic Validation Project (Wise, Peterson, Hoffman, Campbell, & Arabian, 1990).

These basic steps will have a somewhat different form for each of the major samples. The specific features are summarized below.

End-of-Training Validation. The analyses for this sample have two basic parts. The first entails the prediction of each of the five EOT criterion factors with the Experimental Battery. The second part will use the EOT performance factor scores as predictors, which will be validated against the LVI performance factors. Since there is virtually a one-to-one match between EOT and LVI performance scores, the convergent/divergent validities can be calculated both before and after method variance is controlled.

Longitudinal Validation. The basic validation steps for this sample will be as outlined above. Minor differences would result if the CVI performance model is not confirmed for the LVI sample. However, that is not the expectation.

CVII Validation. For CVII the MOS sample sizes are too small to support within-MOS analyses. Further, the available predictors are limited to the ASVAB, and the ABLE for approximately half the total sample. Consequently, the validation analyses for CVII will be carried out on a pooled sample after standardizing criterion scores within MOS. Within these constraints, the basic validities, incremental validities, and optimal equations will be estimated. Similarly, the results of using a priori weights will be compared to the multiple regression results.

Within CVII there is a small sample of approximately 130 individuals for whom we have complete CVI data (i.e., the Trial Predictor Battery and first-tour performance measures). While the sample is too small to support extensive regression analyses, the a priori weights can be used to estimate validities for ASVAB and the Trial Predictor Battery against second-tour performance. Zero order correlations can be calculated between first-tour performance scores and second-tour performance. The sample will also permit a comparison between the concurrent and longitudinal validity of the ABLE on the same sample of people.

Differential Prediction Across Performance Factors

In conjunction with the basic validation steps described above, the analyses for the second contract year will also include an examination of differential prediction across performance factors. The procedures will be analogous to those described in Wise, McHenry, and Campbell (1990). These analyses will be conducted in both the EOT and LVI samples. Because of smaller within-MOS samples and a limited predictor set, they are not appropriate for the CVII sample.

Longer Term Analyses

Beyond the second year of the contract, the Career Force analysis effort will focus on more system-wide objectives. That is, the focus will shift from determining basic validities to optimizing selection/classification procedures for meeting multiple goals (e.g., maximizing aggregate predicted performance versus minimizing attrition) under a variety of constraints (e.g., time, costs, changes in ASVAB). The procedure will be as outlined in the Career Force Research Plan and will be discussed in more detail in later reports.

To aid the optimization analyses, another longer term goal is to more completely model the structural relations among applicant individual differences, training performance, first-tour performance, and second-tour performance. For example, are the relationships of initial ability and temperament with NCO (second-tour) performance totally mediated by their relationships with first-tour performance, or are there significant independent effects? The capability to answer such questions fully will depend on obtaining adequate samples of individuals in the LVII data collection who were also part of the LVI sample. We hope for a successful outcome.

REFERENCES

- Alderman, D. L., Evans, F. R., & Wilder, G. (1981). The validity of written simulation exercises for assessing clinical skills in legal education. Educational and Psychological Measurement, 41, 1115-1126.
- Allen, S. J., & Hubbard, R. (1986). Notes and commentary: Regression equations for the latent roots of random data correlation matrices with unities on the diagonal. Multivariate Behavioral Research, 21, 393-398.
- Borman, W. C. (1974). The rating of individuals in organizations: An alternate approach. Organizational Behavior and Human Performance, 12, 105-244.
- Brogden, H. E. (1946a). An approach to the problem of differential prediction. Psychometrika, 11, 139-154.
- Brogden, H. E. (1946b). On the interpretation of the correlation coefficient as a measure of predictive efficiency. Journal of Educational Psychology, 37, 65-76.
- Brogden, H. E. (1951). Increased efficiency of selection resulting from replacement of a single predictor with several differential predictors. Educational and Psychological Measurement, 11, 173-195.
- Brogden, H. E., & Taylor, E. K. (1950). The dollar criterion--applying the cost accounting concept to criterion construction. Personnel Psychology, 3, 133-154.
- Campbell, J. P. (Ed.) (1987). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1985 fiscal year (ARI Technical Report 746). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A193 343)
- Campbell, J. P. (Ed.) (1987). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1985 fiscal year (ARI Technical Report 792). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A198 856)
- Campbell, J. P. (Ed.) (1989). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1987 fiscal year (ARI Technical Report 862). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A219 046)

- Campbell, J. P. (Ed.) (1991). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1988 fiscal year (ARI Research Note 91-34). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A233 750)
- Campbell, J. P. (1990). An overview of the Army Selection and Classification Project (Project A). Personnel Psychology, 43, 231-239.
- Campbell, J. P., Dunnette, M. D., Lawler, E., & Weick, K. E. (1970). Managerial behavior, performance, and effectiveness. New York: McGraw-Hill.
- Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. Personnel Psychology, 43, 313-333.
- Campbell, J. P., & Zook, L. M. (Eds.) (1991). Improving the selection, classification, and utilization of Army enlisted personnel: Final Report on Project A (ARI Research Report 1597). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A242 921)
- Campbell, R. C. (1985). Scorer training materials (ARI Working Paper RS-WP-85).
- Cronbach, L., & Gleser, G. (1957). Psychological tests and personnel decisions. University of Illinois Press.
- Dawis, R. V., & Lofquist, L. H. (1984). A psychological theory of work adjustment: An individual-difference model and its applications. Minneapolis: University of Minnesota Press.
- DuBois, P. H. (1964). A test-dominated society: China, 1115 B.C.-1950 A.D. Proceedings, ETS Invitational Conference on Testing.
- Eaton, N. K., & Goer, M. H. (Eds.) (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Technical Appendix to the Annual Report (ARI Research Note 83-37). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A137 117)

- Eaton, N. K., Goer, M. H., Harris, J. H., & Zook, L. M. (Eds.) (1985). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1984 fiscal year (ARI Technical Report 660). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A178 944)
- Flanagan, J. C. (1948). The Aviation Psychology Program in the Army Air Forces (Report 1). AAF Aviation Psychology Program Research Reports, U.S. Government Printing Office.
- Flanagan, J. C. (1954). The critical incident technique. Psychological Bulletin, 51, 327-385.
- Guilford, J. P. (1957). A revised structure of intelligence (Report 19). University of Southern California Psychological Laboratory.
- Gunning, R. (1952). The technique of clear writing. New York: McGraw-Hill Book Company.
- Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 18, 133-146.
- Horst, P. (1954). A technique for the development of a differential prediction battery. Psychological Monographs, No. 380.
- Horst, P. (1955). A technique for the development of a multiple absolute prediction battery. Psychological Monographs, No. 390.
- Hough, L. M. (Ed.) (1988). Literature review: Utility of temperament, biodata, and interest assessment for predicting job performance (ARI Research Note 88-02). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A192 109)
- Hough, L. M., Barge, B. N., & Kamp, J. D. (1987). Non-cognitive measures: Pilot testing. In Norman G. Peterson (Ed.), Development and field test of the Trial Battery for Project A (ARI Technical Report 739), pp. 7-1 through 7-48. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A183 575)
- Hugh, L. M., McCloy, R. A., Ashworth, S. D., & Hough, M. M. (1987). Analysis of temperament/biodata, vocational interest, and work environment preference measures: Concurrent validity sample. Working Paper.

- Hough, L. M., McGue, M. K., Houston, J. S., & Pulakos, E. D. (1987). Non-cognitive measures: Field tests. In Norman G. Peterson (Ed.), Development and field test of the trial battery for Project A (ARI Technical Report 739), pp. 8-1 through 8-39. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A184 575)
- Human Resources Research Organization, American Institute for Research, Personnel Decisions Research Institute, & U.S. Army Research Institute. (1983, May). Improving the selection, classification, and utilization of Army enlisted personnel: Project A Research Plan (ARI Research Report 1332). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A129 728)
- Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & U.S. Army Research Institute. (1984, June). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report (ARI Research Report 1347). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A141 087)
- Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & U.S. Army Research Institute. (1984, October). Improving the selection, classification, and utilization of Army enlisted personnel: Appendixes to the Annual Report, 1984 fiscal year (ARI Research Note 85-14). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & U.S. Army Research Institute. (1985, July). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report synopsis, 1984 fiscal year (ARI Research Report 1393). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A173 824)
- Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & U.S. Army Research Institute. (1987). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1985 fiscal year - Supplement to ARI Technical Report 746 (ARI Research Note 87-54). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A188 267)

- Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & U.S. Army Research Institute. (1988). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1986 fiscal year - Supplement to ARI Technical Report 792 (ARI Research Note 88-36). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A196 274)
- Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & U.S. Army Research Institute. (1987). Improving the selection, classification, and utilization of Army enlisted personnel: Annual Report, 1987 fiscal year - Supplement to ARI Technical Report 746 (ARI Research Note 87-54). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A188 267)
- Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & U.S. Army Research Institute. (1990, May). Building and retaining the career force: New procedures for accessioning and assigning Army enlisted personnel--Research Plan (HumRRO RP-PRD-90-11). Alexandria, VA: Human Resources Research Organization.
- Humphreys, L. G., & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. Multivariate Behavioral Research, 10, 193-206.
- Joreskog, K. G., & Sorbom, D. (1986). LISREL VI: Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods. Morresville, IN: Scientific Software.
- Kass, R. A., Mitchell, K., Grafton, F., & Wing, H. (1983). Factor structure of the ASVAB, Forms 8, 9, and 10: 1981 Army applicant sample (ARI Technical Note 581). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A135 660)
- Kuhn, D. B. (1988). The assignment of knowledge test items to functional and cognitive categories (ARI Research Note 88-28). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A196 156)
- Maier, M. H., & Truss, N. R. (1983). Original scaling of ASVAB Forms 5/6/7: What went wrong? (CRC No. 457). Alexandria, VA: Center for Naval Analysis.

- McGuire, C. H., & Babbott, D. (1976). Simulation techniques in the measurement of problem solving skills. Journal of Educational Measurement, 4, 1-10.
- McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. Personnel Psychology, 43, 335-354.
- Montanelli, R. G., & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A Monte Carlo study. Psychometrika, 41, 341-347.
- Motowidlo, S. J., Carter, G. W., & Dunnette, M. D. (1989). A situational inventory for entry-level management positions. In W. C. Borman (Chair), Evaluating "practical IQ": Measurement issues and research applications in personnel selection and performance assessment. Symposium at the Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta.
- Mowry, H. W. (1964). Leadership evaluation and development scale casebook. Los Angeles: Psychological Services.
- Nord, R., & White, L. A. (1988). The measurement and application of performance utility: Some key issues. In B. F. Green, Jr., H. Wing, and A. K. Wigdor (Eds.). Linking military enlistment standards to job performance: Report of a workshop. National Academy Press.
- Nunnally, J. C. (1978). Psychometric theory. New York: McGraw-Hill. p. 249.
- Peterson, N. G. (Ed.) (1987). Development and field test of the Trial Battery for Project A (ARI Technical Report 739). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A184 575)
- Peterson, N. G., Hough, L. M., Dunnette, M. D., Rosse, R. L., Houston, J. S., & Toquam, J. L. (1990). Project A: Specification of the predictor domain and development of new selection/classification tests. Personnel Psychology, 43, 247-276.
- Pulakos, E. D., & Borman, W. C. (Eds.) (1986). Development and field test of Army-wide rating scales and the rater orientation and training program (ARI Technical Report 716). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD B112 857)

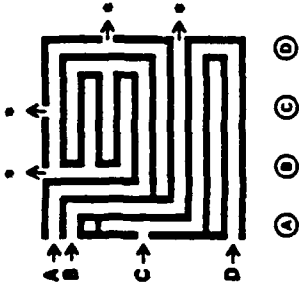
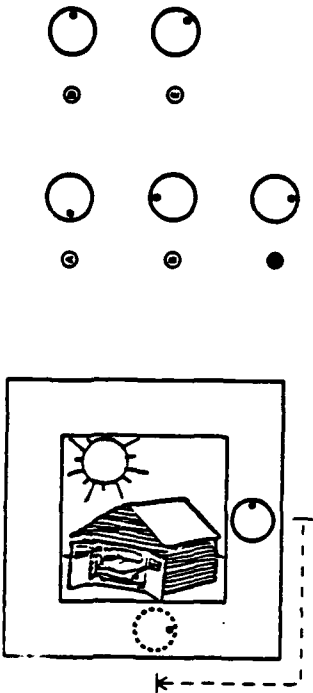
- Rosen, N. A. (1961). How supervise?--1943-1960. Personnel Psychology, 14, 87-99.
- Roznowski, M. A. (1987). Elementary cognitive tasks as measures of intelligence. Doctoral dissertation, University of Illinois at Urbana-Champaign.
- Schmid, J., & Leiman, J. A. (1957). The development of hierarchical factor solutions. Psychometrika, 22, 1, 53-61.
- Smith, I. L. (1983). Use of written simulations in credentialing programs. Professional Practice of Psychology, 4, 21-50.
- Staff, Adjutant General's Office, Personnel Research Branch. (1943a). Personnel research in the Army, I. Background and organization. Psychological Bulletin, 40, 129-135.
- Staff, Adjutant General's Office, Personnel Research Branch. (1943b). Personnel research in the Army, II. The classification system and the place of testing. Psychological Bulletin, 40, 205-211.
- Staff, Adjutant General's Office, Personnel Research Branch. (1943c). Personnel research in the Army, III. Some factors affecting research in the Army. Psychological Bulletin, 40, 237-248.
- Staff, Adjutant General's Office, Personnel Research Branch. (1943d). Personnel research in the Army, IV. The selection of radiotelegraph operators. Psychological Bulletin, 40, 357-371.
- Staff, Adjutant General's Office, Personnel Research Branch. (1943e). Personnel research in the Army, V. The Army specialized training program. Psychological Bulletin, 40, 429-435.
- Staff, Adjutant General's Office, Personnel Research Branch. (1943f). Personnel research in the Army, VI. The selection of truck drivers. Psychological Bulletin, 40, 499-508.
- Stuit, D. B. (1947). Personnel research and test development in the Bureau of Naval Personnel. Princeton University Press.
- Tenopyr, M. L. (1969). The comparative validity of selected leadership scales relative to success in production management. Personnel Psychology, 22, 77-85.

- Thorndike, R. L. (1949). Personnel selection. New York: Wiley.
- Toquam, J. L., Corpe, V. A., & Dunnette, M. D. (1991). Literature review: Cognitive abilities--theory, history, and validity (ARI Research Note 91-28). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A232 638)
- Wise, L. L., McHenry, J. J., & Campbell, J. P. (1990). Identifying optimal predictor composites and testing for generalizability across jobs and across performance factors. Personnel Psychology, 43, 355-366.
- Wise, L. L., Peterson, N. G., Hoffman, R. G., Campbell, J. P., & Arabian, J. M. (1991). Army Synthetic Validation Project: Report of Phase III (Vol. 1) (ARI Technical Report 922). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A235 635)

Appendix A
SAMPLE DESCRIPTIONS OF COGNITIVE MEASURES

DESCRIPTION OF COGNITIVE PAPER-AND-PENCIL MEASURES

Construct/Measure	Description of Test	Sample Item
Spatial Visualization - Rotation		
Assembling Objects	<p>The test contains 36 items with a 18-minute time limit. The subject's task involves figuring out how an object will look when its parts are put back together again. There are two types of problems in the test. In one part, the item shows a picture of labeled parts. By matching the letters, it can be "seen" where the parts should touch when the object is put together correctly. The second type of problem does not label any of the parts. The parts fit together like the pieces of a puzzle. In each section, four possible figures are provided and the subject must pick the correct one.</p>	
Object Rotation	<p>The tests contains 90 items with a 7 1/2-minute time limit. The subject's task involves examining a test object and determining whether the figure represented in each item is the same as the object, only rotated, or is not the same as the test object (e.g., flipped over). For each test object there are five test items, each requiring a response of "same" or "not same."</p>	

Construct/Measure	Description of Test	Sample Item
Spatial Visualization - Scanning		
Maze Test	<p>The test contains 24 items with a 5 1/2-minute time limit. Each item is a rectangular maze with four labeled entrances points and four exit points. The task is to determine which of the four entrances leads to a pathway through the maze and to one of the exit points.</p>	
Spatial Orientation	<p>The test contains 24 items with a 10-minute time limit. Each item contains a picture within a circular or rectangular frame. The bottom of the frame has a circle with a dot inside it. The picture or scene is not in an upright position. The task is to mentally rotate the frame so that the bottom of the frame is positioned at the bottom of the picture. After doing so, the subject must then decide where the dot will appear in the circle.</p>	

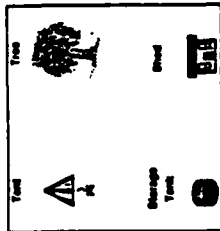
Construct/Measure

Description of Test

Sample Item

Map Test

The test contains 20 items with a 12-minute time limit. Subjects are presented with a map that includes various landmarks such as a barracks, a campsite, a forest, a lake, and so on. Within each item, subjects are provided with compass directions by indicating the direction of one landmark to another, such as "the forest is North of the camp-site." Subjects are also informed of their present location relative to another landmark. Given this information, the subject must determine which direction to go to reach yet another structure or landmark.



1. The shed is due north of the tree. You are at the storage tank. Which direction must you travel to reach the tent?

- ☐ N ☐ NE ☐ E ☐ SE ☐ S ☐ SW ☐ W ☐ NW

2. The tent is due west of the storage tank. You are at the storage tank. Which direction must you travel to reach the tree?

- ☐ N ☐ NE ☐ E ☐ SE ☐ S ☐ SW ☐ W ☐ NW

Induction

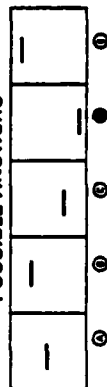
Reasoning Test

The test contains 30 items with a 12-minute time limit. Subjects are presented with a series of four figures. The task is to identify the pattern or relationship among the figures and then to identify from among five possible answers the one figure that appears next in the series.

FIGURE SERIES



POSSIBLE ANSWERS



DESCRIPTION OF COGNITIVE/PERCEPTUAL COMPUTER ADMINISTERED MEASURES

CONSTRUCT/MEASURE

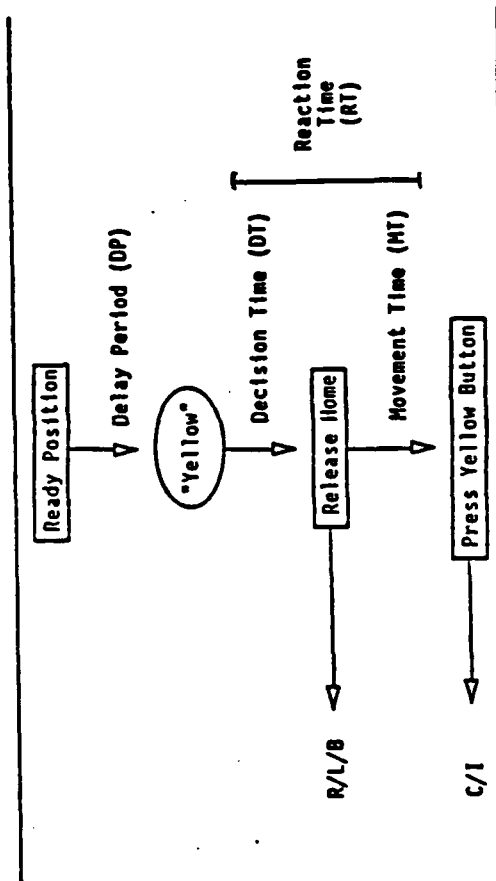
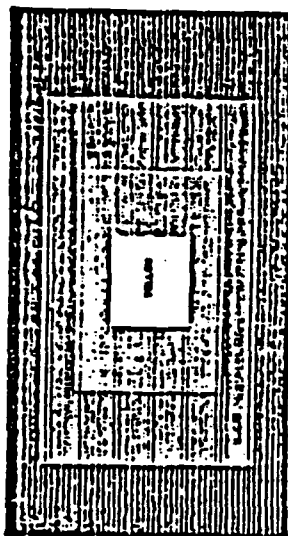
REACTION TIME

Simple Reaction Time

DESCRIPTION OF TEST

The subject is instructed to place his/her hands on the green "home" buttons or in the Ready position. When the subject's hands are in the Ready position, a small box appears on the screen. After a delay period which varies from 1.5 to 3.0 seconds, the word YELLOW appears in the box. At this point, the subject must remove his/her preferred hand from the "home" buttons to strike the yellow key on the testing panel. The subject must then return his/her hands to the ready position to receive the next item. The test contains 15 items. Although it is self-paced, subjects are given 9 seconds to respond before the computer times out and prepares to present the next item.

SAMPLE ITEM



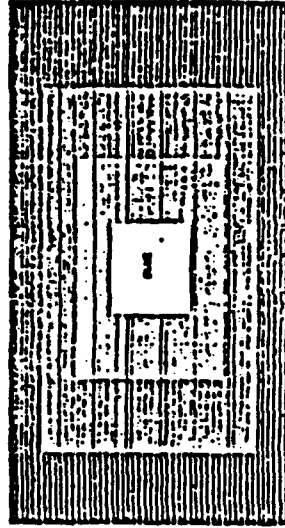
SAMPLE ITEM

DESCRIPTION OF TEST

Choice reaction time is assessed for two response alternatives only. This measure is obtained in virtually the same manner as the simple reaction time measure. The major difference involves stimulus presentation. Rather than presenting the same stimulus (YELLOW) on each trial, the stimulus varies. That is, subjects may see either of the stimuli BLUE or WHITE on the computer screen. When the stimulus appears, the subject is instructed to move his/her preferred hand from the "home" keys to strike the key that corresponds with the term (BLUE or WHITE) appearing on the screen. This test contains 30 items. Although the test is self-paced, the computer is programmed to allow the subject nine seconds to respond before going on to the next item.

CONSTRUCT/MEASURE

Choice Reaction Time

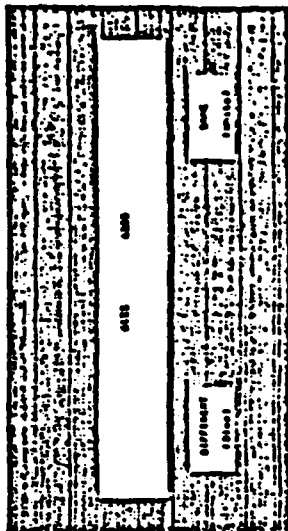


CONSTRUCT/MEASURE
PERCEPTUAL SPEED AND ACCURACY
Perceptual Speed
and Accuracy

DESCRIPTION OF TEST

This test is designed to measure the ability to compare rapidly two visual stimuli presented simultaneously and determine whether they are the same or different. At the beginning of each trial, the subject is instructed to hold down the home keys. After a brief delay, the stimuli are presented. The subject must decide next whether the stimuli are the same or different. He/she must then depress a white button if the stimuli are the same or a blue button if the stimuli are different. Three different "types" of stimuli are used: alpha, numeric, and symbolic. Within each type, the length of the stimulus is varied. Three different levels of length are presented: two-character, five-character, and nine-character. The test consists of 36 trials; the primary dependent variable is the subject's average response time across all trials in which the subject makes a correct response.

SAMPLE ITEM



CONSTRUCT/MEASURE
Target Identification

DESCRIPTION OF TEST

This test was designed to be a job-relevant measure of perceptual speed and accuracy. In this test, the subject is presented with a target object and three stimulus objects. The objects are pictures of military vehicles or aircraft (e.g., tanks, planes, helicopters). The target object is the same as one of the stimulus objects. The target may be rotated in part. The subject must determine which of the three stimulus objects is the same as the target object and then press a button on the response pedestal corresponding to that choice. The test consists of 36 items; the primary dependent variable is the subject's average response time across all trials in which the subject makes a correct response.

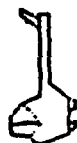
SAMPLE ITEM



TARGET



100%



100%



100%

CONSTRUCT/MEASURE

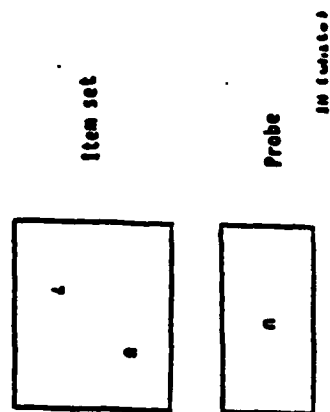
MEMORY

Short Term Memory

DESCRIPTION OF TEST

At the computer console, the subject is instructed to place his/her hands on the green home buttons. The first stimulus set then appears on the screen. A stimulus contains one, three, or five objects (letters or symbols). Following a delay period, the stimulus set disappears. When the probe appears, the subject must decide whether or not it was part of the stimulus set. If the probe was present in the stimulus set, the subject must strike the white key. If the probe was not present, the subject must strike the blue key on the response pedestal. The test includes 36 items. The primary dependent variable is the subject's average response time across those trials in which the subject makes a correct response.

SAMPLE ITEM



CONSTRUCT/MEASURE

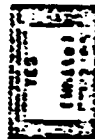
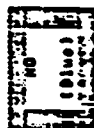
Number Memory

DESCRIPTION OF TEST

At the beginning of each trial of this test, the subject is presented with a single number on the computer screen. After studying the number, the subject is instructed to push a button to receive the next part of the problem. When the subject presses the button, the first part of the problem disappears and another number appears along with an operation term (e.g., "Add 9" or Subtract 6"). Once the subject has combined the first number with the second, he/she must press a button to receive a new number and operation term. This procedure continues until a solution to the problem is presented. The subject must then indicate whether the solution presented is correct or incorrect. In total, the test consists of 28 such items.

SAMPLE ITEM

Start with 14
Divide by 7
Multiply by 8



Target Tracking 1

This is a pursuit tracking test. On each trial of the test, subjects are shown a path consisting entirely of vertical and horizontal line segments. At the beginning of the path is a target box. Centered in the box is a crosshair. As the trial begins, the target starts to move along the path at a constant rate of speed. The subject's task is to keep the crosshair centered within the target at all times. The subject uses a joystick to control movement of the crosshair. The subject's score on this test is the average distance from the center of the crosshair to the center of target across all 18 test trials.

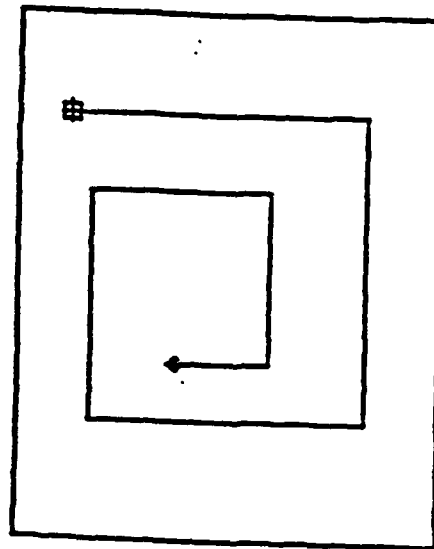
A diagram of a rectangular maze. The maze is defined by a thick black border. Inside, there is a single path that starts from the bottom center, goes up, turns right, then left, then right again, and finally up to a goal symbol (a small circle with a cross) at the top right. A start symbol (a small circle with a cross) is located at the bottom center of the path.

CONSTRUCT/MEASURE
MULTILIMB COORDINATION
Target Tracking 2

DESCRIPTION OF TEST

This is a test of multilimb coordination. The test is virtually identical to Target Tracking #1. The only difference is that the subject must use two sliding resistors (instead of a joystick) to control movement of the crosshair. The first sliding resistor controls movement of the crosshair in the vertical plane, while the second sliding resistor controls movement of the crosshair in the horizontal plane. As with Target Tracking #1, the subject's score on this test is the average distance from the center of the crosshair to the center of the target across all 18 test trials.

SAMPLE ITEM



DESCRIPTION OF PSYCHOMOTOR COMPUTER ADMINISTERED MEASURES

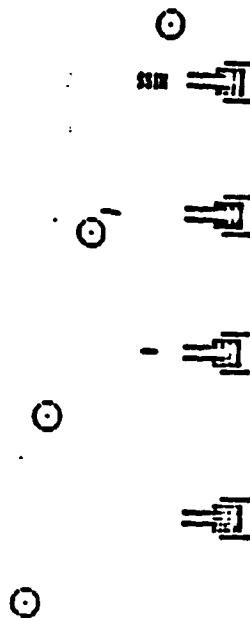
CONSTRUCT/MEASURE
MOVEMENT JUDGMENT

Cannon Shoot

DESCRIPTION OF TEST

At the beginning of each trial of this test, a stationary cannon appears on the computer console. The starting position of this cannon varies from trial to trial (i.e., it is positioned on the top, bottom, or side of the screen). The cannon is capable of firing a shell. The shell travels at a constant speed on each trial. Shortly after the cannon appears, a circular target moves onto the screen. This target moves in a constant direction at a constant rate of speed throughout the trial, though the speed and direction vary from trial to trial. The subject's task is to push a response button to fire the shell such that the shell intersects the target when the target crosses the shell's line of fire. The test includes 36 items. The primary dependent variable is a deviation score indicating the difference between time of fire and optimal fire time (e.g., direct hits yield a deviation score of zero.)

SAMPLE ITEM



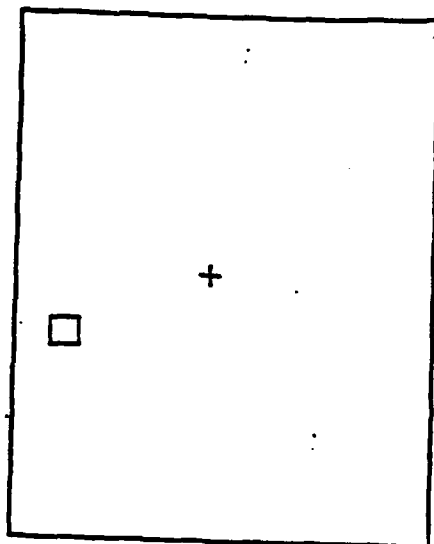
CONSTRUCT/MEASURE
PRECISION/STEADINESS

Target Shoot

DESCRIPTION OF TEST

At the beginning of a trial on this test, a crosshair appears in the center of the screen and a target box appears at some other location on the screen. The target then begins to move about the screen in an unpredictable manner, frequently changing speed and direction. The subject can control movement of the crosshair using a joystick. The subject's task is to move the crosshair into the center of the target. When this has been accomplished, the subject must press a red button on the response pedestal to "fire" at the target. The subject must do this before the time limit on each trial is reached. The subject receives three scores on this test. The first is the percentage of "hits" (i.e., the subject fires at the target when the crosshair is inside the target box). The second is the average time elapsed from the beginning of the trial until the subject fires at the target. The third score is the average distance from the center of the crosshair to the center of the target at the time the subject fires at the target. The test consists of 30 trials.

SAMPLE ITEM



Appendix B
DEFINITIONS OF ELEMENTS IN THE NON-COGNITIVE INVENTORIES

ASSESSMENT OF BACKGROUND AND LIFE EXPERIENCES (ABLE)

Emotional Stability - assesses the amount of emotional stability and tolerance for stress a person possesses. The well-adjusted person is generally calm, displays an even mood, and is not overly distraught by stressful situations. He or she thinks clearly and maintains composure and rationality in situations of actual or perceived stress. The poorly adjusted person is nervous, moody, and easily irritated, tends to worry a lot, and "goes to pieces" in time of stress.

Self-Esteem - is defined as the degree of confidence a person has in his or her abilities. A person with high self-esteem feels largely successful in past undertakings and expects to succeed in future undertakings. A person with low self-esteem feels incapable and is self-doubting.

Cooperativeness - assesses the degree of pleasantness versus unpleasantness a person exhibits in interpersonal relations. The agreeable and likeable person is pleasant, tolerant, tactful, helpful, not defensive and is generally easy to get along with. His or her participation in a group adds cohesiveness rather than friction. A disagreeable and unlikable person is critical, fault-finding, touchy, defensive, alienated, and generally contrary.

Conscientiousness - assesses a person's tendency to be reliable. The person who scores high on this scale is well organized, planful, prefers order, thinks before acting, and holds him- or herself accountable. The person who scores low tends to be careless and disorganized, and acts on the spur of the moment.

Nondelinquency - assesses a person's acceptance of laws and regulations. The person who scores high on this scale is rule abiding, avoids trouble, and is trustworthy and wholesome. The person who scores low on this scale is rebellious, contemptuous of laws and regulations, and neglectful of duty or obligation.

Traditional Values - assesses a person's acceptance of societal values. The person who scores high on this scale accepts and respects authority and the value of discipline. The person who scores low on this scale is unconventional or radical and questions authority and other established norms, beliefs, and values.

Work Orientation - assesses the tendency to strive for competence in one's work. The work-oriented person works hard, sets high standards, tries to do a good job, endorses the work ethic, and concentrates on and persists in the completion of the task at hand. The less achievement-oriented person has little ego involvement in his or her work, does not expend much effort, and does not feel that hard work is desirable.

Internal Control - assesses a person's belief in the amount of control people have over rewards and punishments. The person with an internal locus of control believes that there are consequences associated with behavior and that people control what happens to them by what they do. The person with an external locus of control believes that what happens to people is beyond their personal control.

Energy Level - assesses the amount of energy and enthusiasm a person has. The person high in energy is enthusiastic, active, vital, optimistic, cheerful, zesty, and has the energy to get things done. The person low in energy is lethargic, pessimistic, and tired.

Dominance - is defined as the tendency to seek and enjoy positions of leadership and influence over others. The highly dominant person is forceful and persuasive when adopting such appropriate behavior. The relatively non-dominant person is less inclined to seek leadership positions and is timid about offering opinions, advice, or direction.

Physical Condition - measures the frequency and degree of participation in sports, exercise, and physical activity. Individuals high on this scale actively participate in individual and team sports and/or exercise vigorously several times per week. Those low on this scale have participated only minimally in athletics and exercise infrequently.

Unlikely Virtues - is designed to detect intentional distortion of one's self-description in a favorable direction. High scorers evade answering the ABLE questions frankly and honestly.

Self-Knowledge - consists of items designed to elicit information about how self-aware and introspective the individual is.

Non-Random Response - consists of items that have obvious correct and incorrect response options. The correct options are so obvious that a person responding incorrectly is either inattentive to item content or unable to read or understand the items.

Poor Impressions - measures a variety of negative characteristics. It was developed because of concern that, if the military were to return to a draft, some respondents might distort their self-descriptions in a negative direction to avoid mandatory military service.

ARMY VOCATIONAL INTEREST CAREER EXAMINATION (AVOICE)

Realistic Interests - is defined as a preference for concrete and tangible activities, characteristics, and tasks. Persons with realistic interests enjoy and are skilled in manipulation of tools, machines, and animals, but find social and educational activities and situations aversive. Realistic interests are associated with occupations such as mechanic, engineer, and wildlife conservation officer; negatively associated with such occupations as social work and artist. Scales in the AVOICE that measure realistic interests are: Mechanics, Heavy Construction, Electronics, Electronic Communication, Drafting, Law Enforcement, Fire Protection, Audiographics, Rugged Individualism, Firearms Enthusiast, Combat, and Vehicle Operator.

Conventional Interests - refers to one's degree of preference for well-ordered, systematic, and practical activities and tasks. Persons with conventional interests may be characterized as conforming, unimaginative, efficient, and calm. Conventional interests are associated with occupations such as accountant, clerk, and statistician; negatively associated with occupations such as artist or author. AVOICE scales that measure Conventional

interests are: Clerical/ Administration, Warehousing/Shipping, Food Service--Professional, and Food Service--Employee.

Social and Enterprising Interests - are defined as the amount of liking one has for social, helping, and teaching activities as well as persuasive and leadership activities and tasks. The one AVOICE scale that measures both Social and Enterprising interests is Leadership/Guidance.

Investigative Interests - refers to one's preference for scholarly, intellectual, and scientific activities and tasks. Persons with investigative interests enjoy analytical, ambiguous, and independent tasks, but dislike leadership and persuasive activities. Investigative interests are associated with such occupations as astronomer, biologist, and mathematician; negatively associated with occupations such as salesman or politician. AVOICE scales that measure Investigative interests are Medical Services, Mathematics, Science/Chemical, and Computers.

Artistic Interests - are defined as a person's degree of liking for unstructured, expressive, and ambiguous activities and tasks. Persons with artistic interests may be characterized as intuitive, impulsive, creative and nonconforming. Artistic interests are associated with such occupations as writer, artist, and composer; negatively associated with occupations such as accountant or secretary. The one AVOICE scale that measures Artistic interests is Aesthetics.

JOB ORIENTATION BLANK (JOB)

Job Pride - includes preferences for work environments that are characterized by such positive characteristics as friendly coworkers, fair treatment, and comparable pay. Persons who score high on this scale like the work environment to allow them to feel a sense of accomplishment and to receive recognition for accomplishment.

Job Security/Comfort - includes preferences for work environments that provide secure and steady employment, environments where persons receive good training and can utilize their abilities.

Serving Others - includes preferences for work environments where persons are reinforced for doing things for other people and for serving others through the work performed.

Job Autonomy - includes preferences for work environments that reinforce independence and responsibility. Persons who score high on this construct prefer to work alone, try out their own ideas, and decide for themselves how to get the work done.

Job Routine - includes preferences for work environments that lack variety, where people do the same or similar things every day, have about the same level of responsibility for quite a while, and follow others' directions.

Ambition - measures preferences for work environments that have prestige and status. Persons who score high on this scale prefer work environments that have opportunities for promotion and for supervising or directing others' activities.